Bestimmung der Querschnittsfunktion des menschlichen Stimmkanals aus dem Sprachsignal

Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Georg-August-Universität zu Göttingen

vorgelegt von

Hans Werner Strube

aus Grömitz

Göttingen 1974

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41534922_0020

Bestimmung der Querschnittsfunktion des menschlichen Stimmkanals aus dem Sprachsignal

Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Georg-August-Universität zu Göttingen

vorgelegt von

Hans Werner Strube

aus Grömitz

Göttingen 1974

D 7

UNIV.
LEIDEN
BIBL.

Referent: Prof. M. R. Schroeder

Korreferent: Prof. H. U. Harten

Tag der mündlichen Prüfung: 31. 10. 1974

Zusammenfassung

Die bisherigen Ansätze zur Bestimmung der Querschnittsfunktion des Stimmkanals aus dem Sprachsignal werden dargestellt und auf ihre praktische Brauchbarkeit untersucht. Idealisierte Modellvorstellungen sowie fehlerhafte Messung der relevanten akustischen Größen verfälschen die Ergebnisse evtl. so stark, daß die Anwendung direkter Rechenverfahren theoretisch kaum haltbar erscheint. Die Verwendung von Parameter-Modellen des Stimmkanals erlaubt dagegen komplizierte Modelle und die a-priori-Einschränkung der Menge möglicher Querschnittsfunktionen. Hier werden ein Kosinusreihen-Modell und ein "physiologisches" Artikulatormodell benutzt. Neben der Strahlungsimpedanz wird auch die endliche Wandimpedanz des Stimmkanals berücksichtigt. Die Modellparameter werden iterativ durch Anpassung von akustischen Größen des Modells an die des Sprachsignals bestimmt. Anpassung von Spektren versagt; stattdessen wird ein Stück des Schalldruckverlaufs selbst angepaßt. Dies entspricht einer Art Prädiktion unter Nebenbedingungen. Es erlaubt die Elimination des unbekannten glottalen Quellenspektrums durch Beschränkung der Analyse auf ein Glottisverschlußintervall. Zur Bestimmung von dessen Beginn aus dem Sprachsignal wird ein Verfahren entwickelt.

Beide Modelle liefern mit synthetischen Vokalen recht gute Ergebnisse. Das Artikulatormodell versagt jedoch in unsystematischer Weise, gestattet wenige Freiheitsgrade und ist sehr rechen- und speicherintensiv. Das Kosinusreihen-Modell arbeitet sicherer; die Konvergenz der Iteration wird dabei durch anfängliches Festhalten der Koeffizienten der Kosinusreihe mit geradem Index erheblich verbessert. Ein Test dieses Modells an natürlichen Vokalen liefert vernünftig aussehende Ergebnisse außer bei enger Mundöffnung. Die Schalldämpfung im Stimmkanal müßte wohl ins Modell einbezogen werden.

Inhalt

	Seite
1. Einleitung	1
2. Der Stimmkanal als akustisches System	3
2.1. Vierpoldarstellung	3
2.2. Schallausbreitung im inhomogenen, hartwandigen Rohr	5
2.3. Diskrete Approximation des Übertragungssystems	6
2.3.1. Diskretisierung der Zeit	6
2.3.2. Diskretisierung der Leitung	7
2.4. Der Einfluß der mitschwingenden Wand	11
3. Modelldarstellung des Stimmkanals in dieser Arbeit	13
3.1. Grundlegendes	13
3.2. Die Wahl der Wandadmittanz	15
3.3. Kontinuierliche Längenvariation	17
4. Herstellung und Vorverarbeitung der Sprachproben	19
4.1. Vokalsynthese	19
4.2. Aufnahme natürlicher Vokale	20
4.3. Bestimmung des Glottisverschlußzeitpunktes aus dem Sprachsignal	22
5. Akustische Bestimmung der Querschnittsfunktion	32
5.1. Bisherige Verfahren	32
5.1.1. Verwendung der Eingangsimpedanz oder zweier Sätze von Eigenwerten	32
5.1.2. Verwendung des Sprachsignals	35
5.2. Kritische Untersuchung der Verfahren von Atal und Wakita	38
5.2.1. Die beiden Verfahren	38
5.2.2. Theoretische Überlegungen	39
5.2.3. Modellrechnungen und Experimente	41
5.2.4. Folgerung	52
5.3. Versuche mit iterativer Modellanpassung	52
5.3.1. Grundlegendes	52
5.3.2. Das Iterationsprogramm	54

Seite

5.3.3. Die benutzten Modelle für die Querschnittsfunktion 57
5.3.3.1. Das Kosinusreihen-Modell 57
5.3.3.2. Das Artikulatormodell 57
5.3.4. Anpassung von Spektren (synthetische Vokale) 62
5.3.4.1. mit Kosinusreihen-Modell 62
5.3.4.2. mit Artikulatormodell 64
5.3.5. Prädiktive Anpassung 68
5.3.5.1. Die beiden Verfahren 68
5.3.5.2. Artikulatormodell mit Nennerverfahren (synthetische Vokale) 70
5.3.5.3. Artikulatormodell mit AKF-Verfahren (synthetische Vokale) 80
5.3.5.4. Kosinusreihenmodell (synthetische und natürliche Vokale) 87

5.4. Zusammenfassung und Diskussion 95

Anhang

A. Lineare Prädiktion 99
A.1. Grundlegendes 99
A.2. Das Autokovarianzverfahren 99
A.3. Das Autokorrelationsverfahren 100

B. Das Choleski-Verfahren 101

C. Literaturverzeichnis 103

1. Einleitung

Der menschliche Stimmkanal, d.h. das aus Rachen, Mundhöhle und Nase gebildete akustische Übertragungssystem, bestimmt durch seine veränderbare Form wesentlich die Art des jeweils gesprochenen Lautes. Besonders übersichtlich liegen die Verhältnisse bei (nichtnasalierten) Vokalen, die in dieser Arbeit ausschließlich betrachtet werden sollen. Hier besteht der Stimmkanal nur aus Rachen- und Mundraum und wird durch einen aus der Stimmritze (Glottis) austretenden und von den schwingenden Stimmbändern nahezu periodisch modulierten Luftstrom akustisch breitbandig angeregt. Nach Filterung durch den Stimmkanal erfolgt die Abstrahlung durch die Mundöffnung (etwas auch durch die Kanalwand). Im Frequenzbereich bis 4 kHz sind die Querabmessungen des Stimmkanal-Rohres so klein gegen die Wellenlänge, daß man eindimensionale Schallausbreitung annehmen und höhere Moden vernachlässigen kann. Auch die Krümmung des Rohres möge außer Betracht bleiben. Dann läßt sich eine mittlere Längenkoordinate x im Rohr definieren und an jeder Stelle der Querschnitt S messen; hierdurch ist die "Querschnittsfunktion" S(x) definiert. Die durch Bewegung der Artikulatoren (Kiefer, Zunge, Lippen usw.) veränderbare Querschnittsfunktion bestimmt nun die Filtereigenschaften des Stimmkanals. Seine Übertragungsfunktion weist im wesentlichen nur Pole auf, die sich als Maxima im Leistungsspektrum des abgestrahlten Schallsignals äußern. Diese heißen "Formanten"; die Lage der 2-3 niedrigsten Formanten entscheidet über die Art des wahrgenommenen Vokals. In anderer Betrachtungsweise handelt es sich bei den Formanten um Resonanzen (Eigenschwingungen) des Stimmkanals, in dem sich durch Reflexionen stehende Wellen ausbilden können. Die große Bedeutung der Querschnittsfunktion liegt also darin, daß sie das Bindeglied zwischen den physiologischen Parametern (Artikulatorstellungen) und den akustischen Bestimmungsgrößen der Vokale darstellt (Fig. 1).

Es besteht ein erhebliches Interesse, aus akustischen Daten, am besten aus dem Sprachsignal selbst, auf die Querschnittsfunktion und schließlich auf die Artikulatorstellungen zurückzuschließen. Dies wäre ein gutes Mittel zur Erforschung

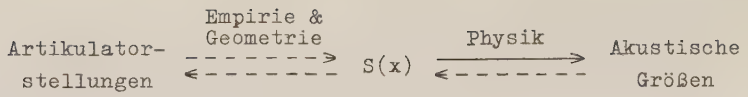


Fig. 1: Die Querschnittsfunktion als Bindeglied
(gestrichelt: nicht genau bekannt)

der Sprechbewegungen, auch zur visuellen Kontrolle des Sprechvorgangs z.B. als Hilfe für Taubstumme [1]. Außerdem lassen sich Parameter, die in engem Zusammenhang zum Artikulationsvorgang stehen, wohl gut zur Codierung von Sprache mit niedriger Bitrate zu Übertragungs- und Speicherzwecken verwenden, auch evtl. zur Analyse und Interpolation. Schon die prädiktive Codierung [2] weist durch ihren engen Zusammenhang mit einer (Pseudo-)Querschnittsfunktion (vgl. 5.2) in diese Richtung. Nichtakustische Methoden zur Ermittlung der Artikulatorstellungen oder ihrer Beziehung zur Querschnittsfunktion, wie Röntgenbilder [3, 4, 5], faseroptische Aufnahmen [6], Ultraschall-Echolot [7], Kieferstellungsmessung [8] usw. sind oft zu langsam zur Bewegungsanalyse, schwer auszuwerten, unangenehm oder gefährlich für die Versuchsperson oder erfassen nur einen Parameter. Solange aber die im einzelnen nur empirisch bestimmbare Beziehung zwischen Artikulationsparametern und Querschnittsfunktion noch nicht einwandfrei bekannt ist, bietet die unabhängige akustische Bestimmung der Querschnittsfunktion zusammen mit den anderen Meßverfahren der Artikulatorstellungen ein Mittel, diese Beziehung zu erforschen. Bisher war eine akustische Prüfung von Annahmen über diese Beziehung nur durch Vergleich von gemessenen und berechneten Formanten möglich (Analysis by Synthesis, entsprechend dem ausgezogenen Pfeil in Fig. 1)[3, 4, 10]. Kenntnis der genannten Beziehung und der Artikulatorbewegungen würde ferner ein Sprachsynthesesystem ermöglichen, dem der Aufbau oder gar die Dynamik des menschlichen Sprechapparates zugrunde liegt und so z.B. Lautübergänge besser wiedergeben kann (vgl. Modelle von [9,10]).

In dieser Arbeit sollen die vorhandenen Ansätze zur direkten akustischen Bestimmung der Querschnittsfunktion, insbesondere aus dem Sprachsignal selbst, dargestellt und auf praktische Brauchbarkeit untersucht werden. Um von der Anfälligkeit

dieser Verfahren gegen Abweichungen von den zugrunde liegenden Idealisierungen ein wenig loszukommen, wird sodann versucht, neue iterative Methoden zur Bestimmung der Querschnittsfunktion, die durch Modelle mit wenigen Parametern dargestellt wird, zu entwickeln. Dabei soll insbesondere die endliche Wandimpedanz des Stimmkanals berücksichtigt und größere Unabhängigkeit von der Glottispulsform erreicht werden.

2. Der Stimmkanal als akustisches System

2.1. Vierpoldarstellung

Der Stimmkanal läßt sich bei eindimensionaler Schallausbreitung als linearer Vierpol betrachten, der von der Glottis gespeist wird und am Mundende mit der Strahlungsimpedanz abgeschlossen ist (Fig. 2)[3, 11, 12]. Dabei wird man in der akustisch-elektrischen Analogie den Schalldruck p mit der Spannung und den Schallfluß u mit dem Strom identifizieren. Ihre Fourier-, Laplace- oder gegebenenfalls z-Transformierten (vgl. 2.3.1) seien P bzw. U. Die Impedanz P/U der Glottis ist komplex mit induktivem Blindanteil und wegen der Stimmbandschwingungen zeitlich schnell variabel, das Gesamtsystem ist also (auch bei fester Stimmkanalform) rheolinear; tatsächlich liegt wegen der Selbsterregtheit der Stimmbandschwingungen sogar Nichtlinearität vor [12, 13, 14]. In der Praxis wird die Glottis aber meist als reine Stromquelle behandelt, da ihre Impedanz (außer für den 1. Formanten) groß gegen die Eingangsimpedanz des Stimmkanals ist [12]. Ihr Einfluß ist aber deutlich meßbar; nach [15] kann der 1. Formant bei offener Glottis bis zu 200 Hz höher liegen als bei geschlossener, auch die

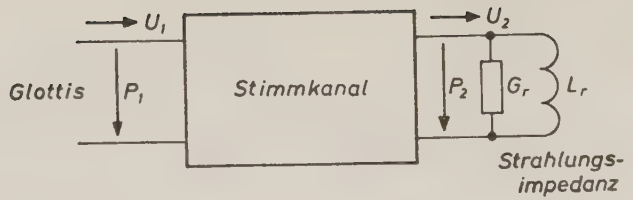


Fig. 2: Der Stimmkanal als Vierpol

Dämpfungen variieren. Das glottale Spektrum ist unregelmäßig und fällt im Mittel mit 12 dB/Oktave ab [16]. - Die Strahlungsimpedanz kann man bei nicht zu hohen Frequenzen gut (für einen Kugelstrahler exakt) als Parallelschaltung einer Induktivität und eines Widerstandes auffassen [11, 17]; auch eine Serienschaltung ist möglich, wobei aber der Widerstand mit dem Quadrat der Frequenz anwächst [11, 12]. Als ganz grobe Näherung genügt oft sogar die Annahme eines Kurzschlusses, wobei der Stimmkanal aber etwas zu verlängern ist[12]. Der Schalldruck im Strahlungsfeld vor dem Mund ist zur Zeitableitung des Schallflusses in der Mundöffnung proportional [18, 19].

Die Stimmkanalform wird i.a. als langsam veränderlich angenommen, so daß man die Theorie zeitlich konstanter Vierpole verwenden kann. Das Verhalten des Systems läßt sich dann gut durch die Kettenmatrix $\underline{A}$ des Vierpols beschreiben, definiert über

$$\begin{pmatrix} P_1 \\ U_1 \end{pmatrix} = \underline{A} \begin{pmatrix} P_2 \\ U_2 \end{pmatrix}, \qquad (2:1)$$

wo der Index 1 die Eingangs-, 2 die Ausgangsgrößen bezeichnet. Die Übertragungsfunktionen für Druck und Schallfluß sowie die Eingangsimpedanzen an den beiden Rohrenden stellen sich für eine Lastadmittanz $Y = U_2/P_2$ mittels der Kettenmatrixelemente folgendermaßen dar:

$$\begin{aligned} H_u &:= U_2/U_1 = Y/(A_{21} + A_{22}Y) \\ H_p &:= P_2/P_1 = 1/(A_{11} + A_{12}Y) \\ Z_1 &:= P_1/U_1 = (A_{11} + A_{12}Y)/(A_{21} + A_{22}Y) = H_u/(H_pY) \\ Z_2 &:= -P_2/U_2|_{U_1=0} = A_{22}/A_{21} \text{ (Glottis geschlossen).} \end{aligned} \qquad (2:2)$$

Die Formantfrequenzen und -bandbreiten entsprechen den (komplexen) Polen der Übertragungsfunktion H_u, die nahe der Fre-

quenzachse liegen und so in gewissen Frequenzbereichen zu einer besonders hohen spektralen Leistungsdichte führen.

Das Innere des Stimmkanal-Vierpols wurde früher z.B. als Serienschaltung zweier Helmholtzresonatoren angesetzt, was aber nur die ersten beiden Formanten erklärt. Da solche "lumped constant"-Modelle außerdem physikalisch nur Sinn haben, wenn der Stimmkanal durch eine Einschnürung deutlich in zwei Hohlräume getrennt ist, setzt man heute den Stimmkanal als inhomogene Leitung an.

2.2. Schallausbreitung im inhomogenen, hartwandigen Rohr

Die die eindimensionale Schallausbreitung beschreibenden Größen sind der über den jeweiligen Querschnitt gemittelte Schalldruck $p(x,t)$ und die über den Querschnitt integrierte Schallschnelle, der Schallfluß $u(x,t)$. Fig. 3 zeigt das Ersatzschaltbild eines infinitesimalen Rohrelements. Die zugehörigen Leitungsgleichungen lauten

$$\begin{aligned} -p' &= L'\dot{u} + R'u \\ -u' &= C'\dot{p} + G'p \end{aligned} \qquad \text{mit} \qquad \begin{aligned} L' &= \rho/S(x) \\ C' &= S(x)/\rho c^2, \end{aligned} \tag{2:3}$$

wo ρ die Luftdichte, c die Schallgeschwindigkeit, $S(x)$ den Rohrquerschnitt bedeuten [3, 11, 12]. Die Verluste R' und G' nehmen mit der Wurzel der Frequenz zu und sind für nicht zu

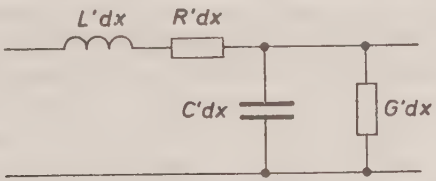


Fig. 3: Ersatzschaltbild eines infinitesimalen Rohrelements

kleines S meist vernachlässigbar. Dann geht aus (2:3) die Webstersche Horngleichung hervor [20]:

$$p'' + \frac{S'}{S}p' = \ddot{p}/c^2, \qquad u'' - \frac{S'}{S}u' = \ddot{u}/c^2 \tag{2:4}$$

oder nach zeitlicher Fouriertransformation:

$$P'' + \frac{S'}{S}P' + (\omega/c)^2 P = 0, \quad U'' - \frac{S'}{S}U' + (\omega/c)^2 U = 0. \quad (2:5)$$

Für ein homogenes Rohr (S' = 0) ist das die gewöhnliche Wellengleichung. Die charakteristische Impedanz des verlustfreien Rohres ist $Z(x) = \rho c/S(x)$, die Phasengeschwindigkeit c. Die Formanten, die Resonanzen des Stimmkanals, entsprechen den Eigenwerten $(\omega/c)^2$ des Sturm-Liouville-Systems (2:5) mit geeigneten Randbedingungen an Glottis- und Mundende [20]. Es gibt theoretisch unendlich viele Eigenwerte; die Beschreibung hat jedoch nur bis etwa 4 kHz physikalischen Sinn, da sich darüber höhere Moden ausbreiten können. Für ein homogenes, an der Glottis hart, am Mund weich abgeschlossenes Rohr von 17 cm Länge liegen die Eigenfrequenzen, die $(n + \frac{1}{2})\lambda$-Resonanzen, bei ca. 0,5, 1,5, 2,5... kHz, etwa entsprechend dem "neutralen Vokal" /ə/.

Voraussetzung für diese Beschreibung ist, daß sich der Querschnitt S als Funktion der Längenkoordinate x nicht zu plötzlich ändert. Anderenfalls werden höhere Moden angeregt [21, 22], die bei tiefen Frequenzen blindgedämpft sind und die sich z.B. für Sprünge von S(x) näherungsweise als Längsinduktivität auswirken. Der Fehler bei Vernachlässigung dieses Effektes ist jedoch gering.

2.3. Diskrete Approximation des Übertragungssystems

2.3.1. Diskretisierung der Zeit [23, 24, 25]

Charakteristisch für die Signalverarbeitung und Simulation von Übertragungssystemen mit einem Digitalrechner ist die Diskretisierung der Zeitachse: Das Signal x(t) wird mit einer Frequenz f_s abgetastet und so in eine Zahlenfolge $x_k := x(k/f_s)$ verwandelt. x(t) ist aus (x_k) genau dann rekonstruierbar, wenn x(t) mit $f_s/2$ frequenzbandbegrenzt ist (Abtasttheorem oder Probensatz). Die Fourier- oder Laplacetransformierte von (x_k), definiert als $\sum_k x_k e^{-i\omega/f_s}$ bzw. $\sum_k x_k e^{-s/f_s}$, ist periodisch mit der Abtastkreisfrequenz, und so kann man durch die Substitution $z = e^{s/f_s}$ zu einer Laurentreihe $X(z) = \sum_k x_k z^{-k}$ übergehen, der "z-Transformierten" $\mathcal{Z}(x_k)$. Der Rand des Einheitskreises in der z-Ebene entspricht der Frequenzachse. Multiplikation mit z^{-n} entspricht einer Zeitverzögerung um n/f_s,

für ganzzahliges n also $\mathfrak{Z}(x_{k-n}) = z^{-n}\mathfrak{Z}(x_k)$. Allgemeiner entspricht der Multiplikation $X(z)Y(z)$ der z-Transformierten eine diskrete Faltung (Convolution) $\sum_n x_n y_{k-n}$ der Zeitfolgen, analog zur Fourier- oder Laplacetransformation. Damit ist die z-Transformation zur Lösung linearer Differenzengleichungen geeignet. Zu simulierende lineare Übertragungssysteme müssen durch solche Gleichungen dargestellt werden. Ist etwa (x_k) das Eingangs- und (y_k) das Ausgangssignal eines Filters, so gilt $\sum_{n=0}^{p} a_n y_{k-n} = \sum_{n=0}^{q} b_n x_{k-n}$, entsprechend $AY = BX$. Die Übertragungsfunktion $H = B/A$ ist immer rational; Stabilität herrscht genau dann, wenn ihre Pole in $|z| < 1$ liegen.

2.3.2. Diskretisierung der Leitung

Um numerische Rechnungen durchführen zu können, wird man die inhomogene Leitung diskretisieren. Eine Möglichkeit ist eine Serienschaltung von Elementen gemäß Fig. 3 oder entsprechenden T- oder Π-Gliedern, die sich digital simulieren oder auch technisch verwirklichen läßt (z.B. LEA, [3]). Eine solche LC-Leiter hat nur endlich viele Formanten und eine obere Grenzfrequenz; der Gesamteinfluß der höheren Formanten muß durch ein Korrekturglied dargestellt werden [3].

Um jedoch Zusammenhänge zwischen Querschnittsfunktion und Übertragungseigenschaften eines inhomogenen Rohres auf einem Digitalrechner auswerten zu können, approximiert man die Gestalt des Rohres zweckmäßiger durch eine Stufenleitung, d.h. durch eine Serienschaltung homogener Segmente gleicher Länge d mit Querschnitten $S_1, \ldots, S_N$ [26, 27]. Wir wollen Verlustlosigkeit voraussetzen. Die das Übertragungsverhalten bestimmenden Größen sind dann die Reflexionsfaktoren der Segmentgrenzen:

$$r_n = \frac{S_{n+1} - S_n}{S_{n+1} + S_n}, \qquad \frac{S_{n+1}}{S_n} = \frac{1 + r_n}{1 - r_n}, \quad |r_n| \leq 1 \qquad (2:6)$$

(hier auf den Schallfluß bezogen; für den Schalldruck wäre das Vorzeichen von r_n umzukehren). Um den Fehler gegenüber der kontinuierlichen Querschnittsfunktion klein zu halten, wird man die Segmentlänge d klein gegen die kleinste betrachtete Wellenlänge und außerdem so klein wählen, daß die Querschnitte

nicht zu stark springen, also die $|r_n| \ll 1$ sind.

Je nach Wahl der akustischen Feldgrößen ergeben sich verschiedene äquivalente Beschreibungsweisen, die unterschiedlich günstig für das jeweils betrachtete Problem sind. Wir betrachten zwei davon (Fig. 4).

1.) P-U-Darstellung

Als Feldgrößen werden Schalldruck P_n und $\rho c \cdot$Schallfluß U_n am Eingang des n-ten Segments gewählt, die Ausgangsgrößen sind P_{N+1}, U_{N+1}. Dann ist die Kettenmatrix $\underline{A}_n$ des Rohres (2:1) zur Beschreibung seines Verhaltens geeignet. Die Kettenmatrix eines homogenen Segments ist [26]

$$\underline{T}_n = \begin{pmatrix} \cos\phi & iS_n^{-1}\sin\phi \\ iS_n \sin\phi & \cos\phi \end{pmatrix}, \qquad \phi := \omega d/c, \tag{2:7}$$

$$\begin{pmatrix} P_n \\ U_n \end{pmatrix} = \underline{T}_n \begin{pmatrix} P_{n+1} \\ U_{n+1} \end{pmatrix}. \tag{2:8}$$

Für die Kettenmatrix des aus den Segmenten 1 bis n gebildeten Rohres gilt

$$\underline{A}_n = \underline{T}_1 \cdots \underline{T}_n = \underline{A}_{n-1}\underline{T}_n. \tag{2:9}$$

Jetzt kann man leicht zu einer diskreten Zeitdarstellung übergehen, indem man als Abtastfrequenz

$$f_s = c/(2d) \tag{2:10}$$

wählt. Da dann $2\phi = \omega/f_s$, erhält man aus (2:7) durch die Er-

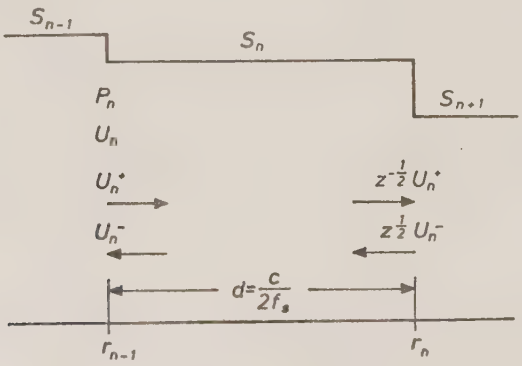


Fig. 4: Segment eines stückweis homogenen Rohres (Stufenleitung)

setzung $z = e^{2i\phi}$ den entsprechenden (exakten!) Ausdruck in z-Darstellung (vgl. 2.3.1):

$$\underline{T}_n = \frac{z^{1/2}}{2}\begin{pmatrix} 1 + z^{-1} & S_n^{-1}(1 - z^{-1}) \\ S_n(1 - z^{-1}) & 1 + z^{-1} \end{pmatrix}. \qquad (2:11)$$

Mit $$z^{\frac{n-1}{2}} P_n(z) = \sum_k p_k^n z^{-k}, \quad z^{\frac{n-1}{2}} U_n(z) = \sum_k u_k^n z^{-k} \qquad (2:12)$$

(Superskript n ist Index, nicht Exponent) liefert der Übergang in den Zeitbereich unter Interpretation von z^{-1} als Verzögerungsoperator aus (2:8) und (2:11) die Differenzengleichungen der Schallausbreitung:

$$\begin{aligned} 2p_k^n &= p_k^{n+1} + p_{k-1}^{n+1} + S_n^{-1}(u_k^{n+1} - u_{k-1}^{n+1}), \\ 2u_k^n &= u_k^{n+1} + u_{k-1}^{n+1} + S_n\,(p_k^{n+1} - p_{k-1}^{n+1}), \end{aligned} \qquad (2:13)$$

und invers:

$$\begin{aligned} 2p_k^{n+1} &= p_k^n + p_{k+1}^n + S_n^{-1}(u_k^n - u_{k+1}^n), \\ 2u_k^{n+1} &= u_k^n + u_{k+1}^n + S_n\,(p_k^n - p_{k+1}^n). \end{aligned} \qquad (2:14)$$

Die Laufzeit durch das Rohr ist dabei schon durch die Zeitverschiebung um $(n-1)/(2f_s)$ in (2:12) berücksichtigt. Setzt man ferner $\underline{A}_n = z^{n/2}\hat{\underline{A}}_n$, so sind die Elemente von $\hat{\underline{A}}_n$ Polynome n-ten Grades in z^{-1}, für deren Koeffizienten ähnliche Rekursionsformeln wie (2:13) gelten. Sie lassen sich jedoch schneller nach einem weiter unten angegebenen Verfahren berechnen.

2.) U^+-U^--Darstellung

In jedem homogenen Segment läßt sich das Schallfeld eindeutig als Überlagerung einer hin- und einer rücklaufenden Welle darstellen [26, 27]:

$$\begin{aligned} P_n &= (U_n^+ + U_n^-)/S_n \\ U_n &= U_n^+ - U_n^- \end{aligned} \quad \text{oder} \quad \begin{pmatrix} P_n \\ U_n \end{pmatrix} = \underline{V}_n \begin{pmatrix} U_n^+ \\ U_n^- \end{pmatrix}, \quad \underline{V}_n := \begin{pmatrix} 1 & 1 \\ S_n^{-1} & S_n^{-1} \end{pmatrix}. \qquad (2:15)$$

Wählt man die Abtastfrequenz gemäß (2:10), so entspricht die Laufzeit durch ein Segment einem Faktor $z^{-1/2}$, und die Stetigkeit von Druck und Schallfluß ergibt (Fig. 4)

$$\begin{aligned} (U_{n+1}^+ + U_{n+1}^-)/S_{n+1} &= (z^{-1/2}U_n^+ + z^{1/2}U_n^-)/S_n\,, \\ U_{n+1}^+ - U_{n+1}^- &= z^{-1/2}U_n^+ - z^{1/2}U_n^-\,, \end{aligned} \qquad (2:16)$$

woraus mit (2:6) folgt

$$\begin{pmatrix} U_n^+ \\ U_n^- \end{pmatrix} = \frac{z^{1/2}}{1 + r_n} \underline{Q}_n \begin{pmatrix} U_{n+1}^+ \\ U_{n+1}^- \end{pmatrix} ; \qquad \underline{Q}_n := \begin{pmatrix} 1 & -r_n \\ -r_n z^{-1} & z^{-1} \end{pmatrix} \tag{2:17}$$

ist die Transmissionsmatrix. Die entsprechende Matrix des n-Segment-Rohres ist

$$\underline{W}_n = \underline{Q}_1 \cdots \underline{Q}_n = \underline{W}_{n-1} \underline{Q}_n \quad ; \quad \begin{pmatrix} U_1^+ \\ U_1^- \end{pmatrix} = z^{n/2} h_n \underline{W}_n \begin{pmatrix} U_{n+1}^+ \\ U_{n+1}^- \end{pmatrix} , \quad h_n := \sum_{k=1}^{n} (1 + r_n)^{-1} . \tag{2:18}$$

Wegen (2:15) bestehen zwischen Ketten- und Transmissionsmatrizen die Beziehungen

$$\underline{T}_n = \frac{z^{1/2}}{1 + r_n} \underline{V}_n \underline{Q}_n \underline{V}_{n+1}^{-1} , \qquad \underline{A}_n = z^{n/2} h_n \underline{V}_1 \underline{W}_n \underline{V}_{n+1}^{-1} , \tag{2:19}$$

wodurch z.B. nochmals (2:11) aus (2:17) folgt. Durch Induktion über n sieht man sofort, daß

$$w_{11}^{(n)}(z) = z^{-n} w_{22}^{(n)}(z^{-1}), \qquad w_{12}^{(n)}(z) = z^{-n} w_{21}^{(n)}(z^{-1}) \tag{2:20}$$

und daß dies Polynome (n-1)ten Grades in z^{-1} sind. Zur rekursiven Berechnung von $\underline{W}_n$ nach (2:18) genügt es wegen (2:20) also z.B.

$$C_n := w_{11}^{(n)} - w_{21}^{(n)}, \qquad D_n := w_{11}^{(n)} + w_{21}^{(n)}, \tag{2:21}$$

beides Polynome n-ten Grades in z^{-1}, zu berechnen. Wir definieren noch für ein solches Polynom $F(z) = \sum_{k=1}^{n} f_k z^{-k}$ das "reziproke" Polynom

$$F^x(z) := z^{-n} F(z^{-1}), \quad f_k^x := f_{n-k}. \tag{2:22}$$

(2:18),(a) geht damit über in

$$\begin{aligned} C_n &= C_{n-1} + r_n z^{-1} C_{n-1}^x \quad \text{oder} & c_k^n &= c_k^{n-1} + r_n c_{n-k}^{n-1} \\ D_n &= D_{n-1} - r_n z^{-1} D_{n-1}^x & d_k^n &= d_k^{n-1} - r_n d_{n-k}^{n-1} \\ C_0 &= D_0 = 1 & c_k^0 &= d_k^0 = \delta_{0k} \end{aligned} \tag{2:23}$$

(die oberen Indizes sind nicht mit Exponenten zu verwechseln!). Dies ist mit (2:20), (2:21) ein einfacher Algorithmus zur Berechnung der Transmissionsmatrix. Wegen (2:15) und (2:19) erhält man auch leicht die Kettenmatrix als

$$\underline{A}_n = \frac{z^{n/2} h_n}{2} \begin{pmatrix} (S_{n+1}/S_1)(D_n + D_n^x) & S_1^{-1}(D_n - D_n^x) \\ S_{n+1}(C_n - C_n^x) & (C_n + C_n^x) \end{pmatrix} . \tag{2:24}$$

Die Abhängigkeit von r_n bzw. S_{n+1} besteht nur scheinbar, da

$h_n(S_{n+1}/S_1)(D_n + D_n^x) = h_{n-1}(S_n/S_1)(D_{n-1} + z^{-1}D_{n-1}^x)$ usw.
Die C_n, D_n können eine einfache physikalische Deutung als inverse Übertragungsfunktionen erhalten: Schreibt man (2:2) speziell für eine konstante reelle Last $Y = S_{n+1}$ hin, so wird nach (2:24) oder auch direkt nach (2:21), (2:18), (2:15)

$$H_u^{-1} = z^{n/2}h_nC_n\ , \qquad H_p^{-1} = z^{n/2}h_n(S_{n+1}/S_1)D_n. \tag{2:25}$$

Die Darstellung der Rohrdämpfung ist in diesem Stufenleitungsmodell schwierig, da dann Phasengeschwindigkeit und charakteristische Impedanzen komplex und frequenzabhängig sind. Die $\sqrt{\omega}$-Abhängigkeit der Verluste läßt auch eine einfache rationale z-Transformationsdarstellung nicht zu.

2.4. Der Einfluß der mitschwingenden Wand

Es gibt experimentelle Hinweise, daß die bisher benutzte Darstellung des Stimmkanals als hartwandiges Rohr mit oder ohne Verluste im Bereich tiefer Frequenzen nicht zutrifft. Wenn man die Formanten eines hartwandigen Rohres berechnet, findet man unterhalb 1 kHz zu geringe Bandbreiten, besonders bei kleiner Lippenöffnung. Wenn die Lippenöffnung Null wird, geht der 1. Formant rechnerisch ebenfalls gegen Null, tatsächlich aber gegen einen endlichen Grenzwert um 150...200 Hz. Diese Diskrepanzen sowie auch die Abhängigkeit der Formanten vom statischen Luftdruck können durch die Annahme erklärt werden, daß die Wand des Stimmkanals nicht starr ist, sondern eine endliche komplexe Impedanz hat [28 bis 32]. Diese wird man als gedämpftes Masse-Feder-System ansetzen, im elektrischen Ersatzbild als LRC-Serienresonanzkreis. Die Kapazität ist so groß, daß sie sich bei Sprachfrequenzen oft vernachlässigen läßt [31]; die Resonanzfrequenz liegt in der Gegend von 30...70 Hz [33].

Eine exakte Lösung der Wellengleichung für gewisse Annahmen über die Verteilung der Wandadmittanz findet man in [34]. Da aber die Wandeffekte nur bei tiefen Frequenzen auftreten, kann man versuchsweise die kontinuierlich verteilte Wandadmittanz durch wenige diskrete RL-Querzweige ersetzen. Betrachtet man nur den ersten Formanten, für den der Stimmka-

nal als Helmholtzresonator approximiert werden kann, so sollte ein einzelner Querzweig genügen. Dann wird das Modell durch Fig. 5 (S. 16) dargestellt. R_w, L_w ist der Querzweig, C das Volumen/ρc^2, L ∝ schwingende Luftmasse, L_r Strahlungslast (fast rein induktiv bei tiefen Frequenzen). Für fast geschlossenen Mund ($L \gg L_w$) hat die Resonanzfrequenz F_1 (1. Formant) einen Grenzwert $f_o = \omega_o/2\pi$. Definiert man ihn als Maximum der Übertragungsfunktion P_2/U_1, so wird

$$\omega_o^2 = (\sqrt{1 + 2/Q^2} - 1/Q^2)/(L_w C), \quad Q := \frac{1}{R_w}\sqrt{\frac{L_w}{C}} . \qquad (2:26)$$

Bei Definition als Imaginärteil des Pols der Übertragungsfunktion wird

$$\omega_o^2 = (1 - Q^{-2}/4)/(L_w C) ; \qquad (2:27)$$

dies ist auch die Frequenz einer gedämpften freien Schwingung. Weiter kann man annehmen, daß für fast geschlossenen Mund die Hauptenergie durch die Wand statt durch den Mund kommt; d.h. ein Teil von L_w ist eine Strahlungslast (der ohmsche Anteil werde vernachlässigt). Dann hat die entsprechende Übertragungsfunktion $H_w := P_{L_w}/U_1$ ein Maximum bei

$$\omega_o^2 = 1/(L_w C). \qquad (2:28)$$

Für nicht zu kleines Q sind die drei Frequenzen sehr ähnlich.

Bei offenem Mund und fehlendem Querzweig ist die Formantfrequenz

$$F_{1a} = 1/(2\pi\sqrt{LC}). \qquad (2:29)$$

Durch den Querzweig wird die Frequenz erhöht; mit $Q^2 \gg 1$ gilt

$$F_1^2 = F_{1a}^2 + f_o^2 \qquad [30]. \qquad (2:30)$$

Das gemessene $f_o \approx 150$ Hz entspricht einem L_w der Größenordnung 0,02 g cm^{-4}. Ist b_o die Bandbreite von f_o, d.h. (für $Q^2 \gg 1$) $b_o/f_o = 1/Q$, so wird die Bandbreite des ersten Formanten ohne sonstige Dämpfungsquellen [30]

$$B_1 = b_o/(1 + L_w/L) = b_o(f_o/F_1)^2. \qquad (2:31)$$

b_o hat die Größenordnung 100 Hz.

Da der Einfluß auf die höheren Formanten weniger wichtig ist, kann zur rechnerisch einfachsten Simulation der Wandadmittanz ein einziger RL-Querzweig am Eingang der Stimmkanal-

Leitung angesetzt werden; tatsächlich scheint die Stimmkanalwand dort auch am weichsten zu sein [32]. In der Kettenmatrix der Leitung ist dann lediglich die Ersetzung

$$A_{2i} \longmapsto A_{2i} + Y_w A_{1i} \ , \quad i = 1, 2, \qquad (2:32)$$

durchzuführen, wo $Y_w = 1/(R_w + sL_w)$ die Queradmittanz bedeutet. Für eine genauere Darstellung könnte man diskrete Leitungen nach 2.3.2 mit je einem Querzweig in jedem Segment ansetzen; dies würde die rechnerische Simulation aber wesentlich komplizieren.

3. Modelldarstellung des Stimmkanals in dieser Arbeit

3.1. Grundlegendes

Der Stimmkanal wird als verlustfreie Stufenleitung modelliert (2.3.2). Die Segmentlänge beträgt 0,5 cm, was nach (2:10) einer Abtastfrequenz f_s von 35,3 kHz entspricht (die Schallgeschwindigkeit bei Körpertemperatur beträgt 353 m/s). In allen Fällen wird nicht die Schallausbreitung selbst simuliert, sondern über die Kettenmatrix die Übertragungsfunktion berechnet. Die endliche Wandimpedanz des Stimmkanals wird durch einen RL-Querzweig am Glottisende im Sinne der in 2.4 erwähnten Näherung dargestellt. Hat dieser die Admittanz Y_w und ist Y_r die Strahlungsadmittanz, $\underline{A}$ die nach (2:24) berechnete Kettenmatrix des hartwandigen Rohres, so wird nach (2:2) und (2:32) die Schallfluß-Übertragungsfunktion

$$H_u = Y_r/((A_{21} + Y_w A_{11}) + (A_{22} + Y_w A_{12})Y_r)$$
$$= Y_r/((A_{21} + Y_r A_{22}) + (A_{11} + Y_r A_{12})Y_w). \qquad (3:1)$$

Die Strahlungsadmittanz wird als Parallelschaltung einer Induktivität L_r und eines Leitwerts G_r angesetzt, wobei die Mundöffnung als kreisförmiger Kolbenstrahler in einer ebenen Wand aufgefaßt wird [11, 12]; solange die Mundöffnung klein gegen die Wellenlänge ist, spielt ja die genaue Form der Öffnung eine geringe Rolle. Ist S_N der Querschnitt des N-ten Segments, d.h. des Mundes, so wird

$$Y_r = G_r + 1/(sL_r), \quad 1/L_r = \frac{3}{8}c\pi^{3/2}\sqrt{S_N} = 73710 \text{ cm/s}\sqrt{S_N},$$
$$G_r = \frac{9\pi^2}{128} S_N = 0{,}69396\ S_N, \qquad (3:2)$$

in Einheiten 1/ρc. Alternativen wären z.B. Kugelstrahler oder Kolbenstrahler in einer harten Kugel, was die Faktoren in diesen Formeln geringfügig verändern würde. Um einen entsprechenden Ausdruck für Y_r in diskreter Zeitdarstellung zu finden, muß eine z-Transformation 2. Ordnung verwendet werden ("bilineare Transformation", $s \mapsto 2f_s(1 - z^{-1})/(1 + z^{-1})$ [23, 24, 35]); Transformationen 1. Ordnung liefern eine zusätzliche Dämpfung, die auch für beliebig kleine Frequenzen einen positiven Grenzwert behält. Es wird damit

$$Y_r = \frac{g_1 - g_2 z^{-1}}{1 - z^{-1}}, \quad g_1 = G_r + \frac{1}{2f_s L_r}, \quad g_2 = G_r - \frac{1}{2f_s L_r}. \qquad (3:3)$$

Die bilineare Darstellung von Y_w in (2:32) ist

$$Y_w = \frac{1 + z^{-1}}{w_1 - w_2 z^{-1}}, \quad w_1 = 2f_s L_w + R_w, \quad w_2 = 2f_s L_w - R_w. \qquad (3:4)$$

Die Frequenzachsenverzerrung bei der bilinearen Transformation spielt in dieser Arbeit keine Rolle, da die Abtastfrequenz so hoch ist (35,3 kHz). Die gesamte Übertragungsfunktion (3:1) lautet nun mit (3:3), (3:4):

$$H_u(z) = \frac{(g_1 - g_2 z^{-1})(w_1 - w_2 z^{-1})}{(A_{21}+g_1 A_{22}-z^{-1}(A_{21}+g_2 A_{22}))(w_1 - w_2 z^{-1}) + (A_{11}+g_1 A_{12}-z^{-1}(A_{11}+g_2 A_{12}))(1 + z^{-1})}. \qquad (3:5)$$

Der Zähler ist ein Polynom in z^{-1} vom 2. Grad, der Nenner eines vom (N+2)ten Grad, wenn man einen Laufzeitfaktor $z^{N/2}$ herauszieht (N = Zahl der Segmente).

3.2. Die Wahl der Wandadmittanz

Um die Werte R_w und L_w für den die Wandadmittanz darstellenden Querzweig realistisch zu wählen, mußten einige Kontrollrechnungen durchgeführt werden. Nach 2.4 scheint $L_w = 0{,}02$ g cm^{-4} ein brauchbarer Wert zu sein. R_w wurde auf 25 g $cm^{-4}s^{-1}$ geschätzt. Dann bewirkt der Querzweig bei einem Rohr von 17 1-cm-Segmenten ($S_1=\ldots=S_{16} = 5$ cm^2, S_{17} variabel) Verschiebungen des 1. Formanten nach oben, die nach (2:30) einem Grenzwert f_o von 133 bis 184 Hz für $0{,}2\ cm^2 \leq S_{17} \leq 5\ cm^2$ entsprechen. Daß der Wert für f_o mit größerer Mundöffnung zunimmt, liegt daran, daß die Druckverteilung dann nicht wie beim Helmholtzresonatormodell konstant ist, sondern ein Maximum an der Glottis hat, wodurch der Einfluß des diskreten Querzweiges sich erhöht. Wäre der Querzweig kontinuierlich verteilt, so sollten etwas kleinere Werte von L_w, R_w ausreichen. Für praktisch geschlossenen Mund ($S_{17} = 10^{-6} cm^2$) wurden die Resonanzkurven für die Übertragung durch Mund und Wand berechnet. $|P_2/U_1|^2$ erhält man aus H_u, dabei ist (3:5) für verschiedene Frequenzen unter Beachtung von $z = e^{2\pi i f/f_s}$ auszuwerten. $|P_{L_w}/U_1|^2$ folgt daraus durch Multiplikation mit $\frac{(\omega L_w)^2}{R_w^2 + (\omega L_w)^2}$, wie Fig. 5 nahelegt. Fig. 6 zeigt die Kurven für (a) $R_w = 25$, $L_w = 0{,}02$ und (b) $R_w = 15$, $L_w = 0{,}015$. Der erste Fall scheint eine etwas zu große Bandbreite zu liefern. Für dieselben Werte wurden die Formantfrequenzen und -bandbreiten für Querschnittsfunktionen nach Fant [3] berechnet; die Segmentzahl beträgt 17 bis 20, d = 1 cm; die S_n sind Mittelwerte zweier benachbarter Querschnittswerte nach Fant, die dort alle 0,5 cm gegeben sind. In Fig. 7 ist B_1 über F_1 aufgetragen. Man erhält die gleiche V-Form mit einem Minimum nahe 500 Hz wie in den Messungen nach [31], Fig. I.-9; dort sind die Bandbreiten etwas größer, da in den Modellrechnungen Verluste durch Viskosität und Wärmeleitung nicht berücksichtigt wurden (vgl. auch [34]). Die Fälle (a) und (b) wurden nicht unterschieden; denn die Dämpfung hängt fast nur von R_w/L_w^2 ab, was in beiden Fällen fast gleich ist. Allein aus Bandbreitenmessungen lassen sich R_w und L_w also nicht bestimmen, man muß zusätzlich die Grenzfrequenz f_o kennen. In allen weiteren Ex-

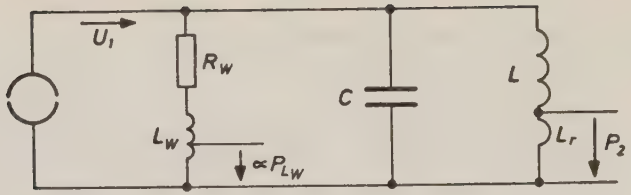


Fig. 5: Helmholtzresonatormodell für den 1. Formanten bei mitschwingender Wand; R_w, L_w: Wandadmittanz

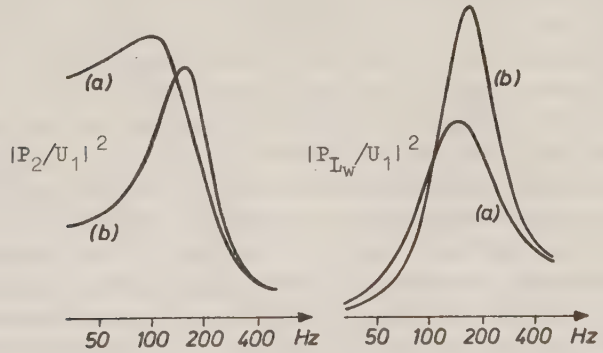


Fig. 6: Resonanzkurven des 1. Formanten für Übertragung durch Mund (links) und Wand (rechts) bei fast verschwindender Mundöffnung; homogenes Rohr 16 cm x 5 cm^2;
(a) R_w = 25 g $cm^{-4}s^{-1}$, L_w = 0,02 g cm^{-4},
(b) R_w = 15 g $cm^{-4}s^{-1}$, L_w = 0,015 g cm^{-4},
Querzweig R_w, L_w am Rohreingang

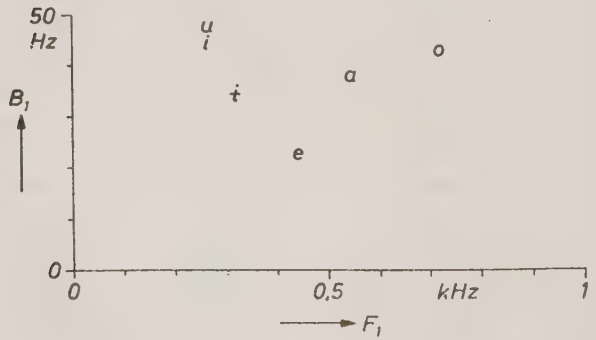


Fig. 7: 1. Formant für Fant-Querschnittsfunktionen mit RL-Querzweig am Rohreingang; R_w, L_w wie in Fig. 6 (a)

perimenten wurden die Werte nach (a) benutzt; die Werte scheinen nicht sehr kritisch zu sein.

3.3. Kontinuierliche Längenvariation

Bei den Experimenten mit dem Artikulatormodell (s.u., 5.3.3.2) muß die Länge des Stimmkanals kontinuierlich veränderbar sein. Das Problem dabei ist, daß die benutzte Darstellung für eine feste Abtastfrequenz f_s bzw. Segmentlänge d nur diskrete Rohrlängen zuläßt: $L = Nd$, $N = 1, 2, 3...$ Für beliebige positive Längenwerte nun gibt es ein ganzzahliges N und ein $\alpha \in [0,1)$ oder $(0,1]$, so daß $L = (N+\alpha)d$. Wir können dann ein homogenes Segment mit Länge αd und Querschnitt S an irgendeiner gewünschten Stelle in das Rohr einsetzen. Seine Kettenmatrix ist offenbar durch (2:7) gegeben, wo S_n und ϕ durch S und $\alpha\phi$ zu ersetzen sind, und formal können wir die z-Transformierte herleiten (ein gemeinsamer Faktor $z^{\alpha/2}$ werde weggelassen):

$$\underline{T} = \frac{1}{2}\begin{pmatrix} 1 + z^{-\alpha} & S^{-1}(1 - z^{-\alpha}) \\ S(1 - z^{-\alpha}) & 1 + z^{-\alpha} \end{pmatrix}. \qquad (3:6)$$

Jetzt sind jedoch für $\alpha \neq 0$ oder 1 die Matrixelemente keine Polynome mehr, so wie auch die aus Kettenmatrix und Lastimpedanz hergeleiteten Übertragungsfunktionen und Impedanzen keine rationalen Funktionen mehr sind, was nach 2.3.1 erforderlich ist. Deshalb muß eine passende Approximation von $z^{-\alpha}$ durch ein Polynom gefunden werden. Der Fehler sollte klein sein besonders bei tiefen Frequenzen, d.h. bei $z^{-1} \approx 1$, so daß eine Entwicklung nach Potenzen von $(z^{-1} - 1)$ geeignet erscheint. In 1. Näherung hat man

$$z^{-\alpha} \approx 1 + \alpha(z^{-1} - 1) = (1 - \alpha) + \alpha z^{-1}. \qquad (3:7)$$

Dies ist als lineare Interpolation zwischen den Funktionen z^0 und z^{-1} auffaßbar. Die Kettenmatrix (3:6) wird zu

$$\underline{T} = \frac{1}{2}\begin{pmatrix} (2-\alpha) + \alpha z^{-1} & S^{-1}\alpha(1 - z^{-1}) \\ S\alpha(1 - z^{-1}) & (2-\alpha) + \alpha z^{-1} \end{pmatrix}. \qquad (3:8)$$

Das Verfahren wurde geprüft durch Berechnung der Resonanzen (Formanten) eines am Ende kurzgeschlossenen homogenen Rohres. Man erhält sie als die Lösungen der Gleichung

$$1 + z^{-(N+\alpha)} = 0; \qquad (3:9)$$

die Formantfrequenzen sind $F_n = (n - \frac{1}{2})f_s/(N + \alpha)$, $n = 1,2,3...$, die Formantbandbreiten B_n sind Null. Jetzt wird (3:9) durch die Polynomgleichung

$$1 + (1 - \alpha)z^{-N} + \alpha z^{-(N+1)} = 0 \qquad (3:10)$$

approximiert. Der resultierende Fehler in der Formantfrequenz ist Null bei $\alpha = 0$, $\alpha \approx 1/2$, $\alpha = 1$ und durchweg klein (von 3. Ordnung) für Frequenzen unterhalb $f_s/4$. Jedoch tritt eine zusätzliche Dämpfung auf, die etwa proportional zu $\alpha(1-\alpha)$ ist. Die maximale zusätzliche Formantdämpfung ist näherungsweise

$$\Delta B_{max} \approx \frac{\pi c}{4L}(F/f_s)^2 = \frac{\pi L}{c}\left(\frac{F}{N+\alpha}\right)^2, \quad F < f_s/4, \qquad (3:11)$$

wo F die Formantfrequenz bezeichnet. Dies erhält man folgendermaßen. Sei $\phi = \pi F/f_s$, $\delta = \pi \Delta B/f_s$, $z = e^{2i\phi-\delta}$. Da nach (3:9) $e^{-2Ni\phi} = -e^{2\alpha i\phi}$ ist, liefert Entwicklung von (3:10) in ϕ bis zur 2. Ordnung

$$0 = 1 - (1-\alpha)(1+2\alpha i\phi-2\alpha^2\phi^2+N\delta) - \alpha(1+2(\alpha-1)i\phi-2(\alpha-1)^2\phi^2+(N+1)\delta)$$
$$= 2\alpha(1-\alpha)\phi^2 - (N+\alpha)\delta, \qquad (3:12)$$

und (3:11) folgt unmittelbar. (3:11) wurde auch numerisch verifiziert. Für ein Stimmkanalmodell mit d = 0,5 cm, N = 34, f_s = 35,3 kHz sind die maximalen zusätzlichen Bandbreiten der ersten vier Formanten 0,3, 3, 8 bzw. 17 Hz, was für die meisten Anwendungen erträglich erscheint, da die Bandbreiten selbst üblicherweise viel größer sind. Wenn weniger Segmente verwendet werden, kann die lineare Näherung zu ungenau werden. Dann ist eine Näherung 3. Ordnung angebracht [36]; sie wurde in dieser Arbeit nicht benutzt.

Bei so kleiner Segmentlänge führt auch die Annahme diskreter Längenwerte nur zu einem Fehler der Formantfrequenzen von wenigen Prozent. Daher ist der Hauptvorteil des Verfahrens nicht eine größere Genauigkeit der Formantfrequenzen, sondern die Fähigkeit zur Darstellung stetiger (bzgl. der Frequenz sogar fast glatter) Formantübergänge bei kontinuierlicher Änderung der Rohrlänge.

In der Praxis wurde das zusätzliche Segment am Mundende angebracht und $\alpha \in (0,1]$ gewählt. Ist dann N die gesamte Segmentzahl (einschließlich des kürzeren), so ist der Nenner der Übertragungsfunktion (3:5) wieder vom Grad N + 2.

4. Herstellung und Vorverarbeitung der Sprachproben

4.1. Vokalsynthese

Um Verfahren zur Bestimmung der Querschnittsfunktion testen zu können, mußten einige synthetische Vokale nach bekannten Querschnittsfunktionen erzeugt werden. Dazu wurde ein angenommenes Eingangssignal durch ein simuliertes Filter mit der aus der Querschnittsfunktion bestimmten Übertragungsfunktion (3:5) geschickt. Das Eingangssignal entspricht der Zeitableitung des glottalen Schallflusses u(t), hierdurch wird der Abstrahlvorgang berücksichtigt (vgl. 2.1), und das Ausgangssignal entspricht dem Schalldruck im Strahlungsfeld. Der glottale Schallfluß wurde gemäß

$$u(t) = \begin{cases} \sin^2 \frac{\pi t}{2T_1} & , \quad 0 \leq t \leq T_1 \\ \sin \frac{\pi(T_2 - t)}{2(T_2 - T_1)} & , \quad T_1 \leq t \leq T_2 \\ 0 & , \quad T_2 \leq t \leq T \end{cases} \qquad (4:1)$$

angesetzt [37] mit T_1 = 2,89 ms (102 Abtastwerte), $T_2 - T_1$ = 1,16 ms (41 Abtastwerte), Periode T = 7,25 ms (256 Abtastwerte), siehe Fig. 8. T_1 und T_2 sind willkürlich, aber "gut aussehend" gewählt; f_s betrug wieder 35,3 kHz; die Periode ist eine Zweierpotenz, um etwa nötige Fouriertransformationen zu erleichtern. Die Pulse werden noch mit einem nahezu phasenlinearen Tiefpaß auf etwa 4,4 kHz bandbegrenzt, dessen Impulsantwort durch

$$h_{k+24} = \begin{cases} (0{,}54 + 0{,}46\cos(\pi k/24))\frac{\sin(13\pi k/48)}{48\sin(\pi k/48)}, & k = \begin{cases} -24,..,-1 \\ 1,..,23 \end{cases} \\ 13/48 \quad , & k = 0 \\ 0 \quad , & \text{sonst} \end{cases} \qquad (4:2)$$

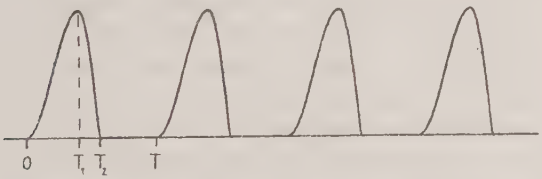


Fig. 8: Glottaler Schallfluß nach (4:1) zur Vokalsynthese

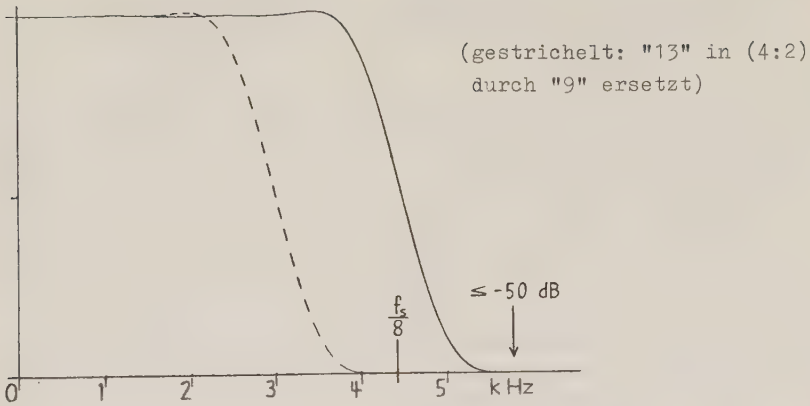


Fig. 9: Betragsquadrat der Tiefpaß-Übertragungsfunktion nach (4:2)

gegeben ist (Fig. 9). Hierdurch weicht das Eingangssignal auch in den Glottisverschlußintervallen etwas von Null ab, aber die Werte sind klein.

Für die Vokalsynthese wurden hauptsächlich die Querschnittsfunktionen nach Fant [3] zugrundegelegt, manchmal auch andere Modelle (siehe 5.3.5.2). Die ersten fünf Perioden des Schalldrucks wurden nicht berücksichtigt, damit Stationarität erreicht werden konnte. Durch Differenzbildung benachbarter Abtastwerte wurde noch einmal eine Höhenanhebung um 6 dB/Okt. zum Ausgleich des abfallenden Glottisspektrums vorgenommen.

4.2. Aufnahme natürlicher Vokale

Bei der Aufnahme der Vokalproben war darauf zu achten, daß die Apparatur im interessierenden Frequenzbereich bis 4,4 kHz neben einem konstanten Amplitudengang auch einen möglichst linearen Phasengang, d.h. eine frequenzunabhängige Zeitverzögerung, aufwies. Denn bei vielen Experimenten sollte der Umstand ausgenutzt werden, daß in einem Teilintervall der Glottisperiode der Eingangsschallfluß des Stimmkanals nahezu verschwindet, da die Glottis geschlossen ist. Eine Veränderung der relativen Amplituden und Phasen der verschiedenen Frequenzkomponenten würde auch im Glottisverschlußintervall das

Eingangssignal von Null verschieden machen. Um solche linearen sowie nichtlineare Verzerrungen und Verschlechterung des Rauschabstandes zu vermeiden, wurde auf jede Analog-Signal-Aufzeichnung verzichtet. Zur Aufnahme wurde ein hochwertiges 1/2-Zoll-Kondensatormikrophon (Brüel & Kjaer 4134) mit einem Zwischenring eigener Konstruktion auf den 1/4-Zoll-FET-Vorverstärker 2618 geschraubt; bei Besprechung unter 90° ist der Frequenzgang nach den mitgelieferten Eichkurven bis 10 kHz völlig flach. Die Kleinheit des Mikrophons hält Beugungseffekte unter 4,4 kHz sehr klein. Auf den Vorverstärker folgte die Vorstufe des bis 200 kHz linearen Meßverstärkers B&K 2608; an seinen (nicht gleichspannungsfreien) Filterausgang wurde über ein RC-Differenzierglied (1 nF, 1.1 kOhm) ein Krohn & Hite-Zweifachfilter 3343 angeschlossen. Das Differenzierglied erfüllt mehrere Aufgaben: Fernhalten der Gleichspannung; Beseitigung tiefstfrequenter Störungen, die reichlich auftraten; Höhenanhebung mit 6 dB/Okt. zum Ausgleich des Sprachspektrums. Die Grenzfrequenz von 14,5 kHz liegt so hoch, damit der Frequenzgang im interessierenden Bereich noch einigermaßen linear ist (Amplitudenabnahme bei 4,4 kHz etwa 4,5 %). Die beiden Stufen des Krohn & Hite-Filters waren als Butterworth-Tiefpässe mit 17 und 7 kHz Grenzfrequenz (je 48 dB/Okt. Abfall) geschaltet, jede verstärkte außerdem zehnfach. Die Grenzfrequenz 17 kHz lag so hoch, da diese Stufe im wesentlichen nur als Verstärker mit möglichst linearem Phasengang arbeiten sollte. Die Abweichung des 7-kHz-Filters vom linearen Phasengang dürfte nach den Kurven im Manual bis 4,4 kHz 10 bis 15° nicht überschreiten. Die Rauschspannung am Filterausgang betrug bei eingeschaltetem Mikrophon wenige mV_{eff}. Von dem kapazitiv stark belastbaren 50-Ohm-Filterausgang führte ein 20 m langes 50-Ohm-Kabel zu einem 12-bit-Analog-Digital-Wandler mit Sample-and-hold-Verstärker, der das Signal mit 35,3 kHz abtastete. Sein Aussteuerbereich von 7 V_{ss} wurde zu mindestens 25 %, nach Möglichkeit sogar voll ausgenutzt. Die Daten wurden on-line über den Direct-Read-Eingang eines Digitalrechners Honeywell H632 eingelesen, in Gleitkommazahlen umgewandelt und mit einer Kennung auf Magnetband geschrieben. Eine weiter Herabsetzung der oberen Grenzfrequenz von 7 auf $\leq$ 4,4 kHz erfolgte erst bei Wiedergebrauch der Daten durch

rechnerische Tiefpaßfilterung mit dem Filter nach Fig. 9.

Die aufzunehmenden Vokale wurden in einem reflexionsfreien Raum aus 20 bis 30 cm Abstand in das Mikrophon gesprochen. Der AD-Wandler konnte von dort fernausgelöst und Bereitschaft und Übersteuerung angezeigt werden. Die übertragenen Abschnitte hatten eine Länge von 1229 Abtastwerten oder 34,8 ms.

4.3. Bestimmung des Glottisverschlußzeitpunktes aus dem Sprachsignal

(In diesem Kapitel wurde die Abtastfrequenz durchweg nach entsprechender Tiefpaßfilterung auf $f_s/4 = 8825$ Hz herabgesetzt, damit der interessierende Frequenzbereich gerade bis zur halben Abtastfrequenz reicht.)

Um den Einfluß der Glottispulsform und der variablen Glottisimpedanz zu verringern, werden in dieser Arbeit Analyseverfahren benutzt, die das Signal nur auf den Glottisverschlußintervallen betrachten. Dazu ist es nötig, die Lage dieser Intervalle zu bestimmen, nach Möglichkeit aus dem Sprachsignal selbst. Zur akustischen Ermittlung des Glottisverschlußzeitpunktes kann man davon ausgehen, daß der Stimmkanal bei nicht zu tiefen Frequenzen in diesem Augenblick am stärksten angeregt wird (die Glottis schließt sich abrupt und öffnet sich langsamer)[14, 16, 37] und daß danach ein Zeitintervall freier Schwingung folgt. Der gesuchte Zeitpunkt wird also in der Nähe der stärksten Amplitudenzunahme des (höhenangehobenen) Schalldrucks liegen, und bei Anwendung der linearen Prädiktion (Anhang A) wird der Prädiktionsfehler dort groß sein, da ja für gute Prädizierbarkeit das Signal als freier Ausschwingvorgang eines Nur-Pole-Filters nach Verschwinden des Eingangssignals darstellbar sein muß. Atal [26] bestimmte "Anregungsimpulse" aus den Maxima des Prädiktionsfehlers. Möglicherweise fallen diese Maxima aber nicht mit den Verschlußzeitpunkten zusammen, da Atal die Glottispulsform durch zwei Pole der Übertragungsfunktion darstellt und Deltaimpuls-Erregung annimmt. Auch ist nicht klar, für welche Intervalle hier vor Kenntnis der Lage dieser Impulse der Prädiktor berechnet wird. Nimmt man ein 30-ms-Segment mit Hamming-Wichtung [38], so zeigt der Prädiktionsfehler an den Stellen starker Anregung meist deutliche Spitzen (Fig. 10, Mitte), die aber nicht immer eindeutig de-

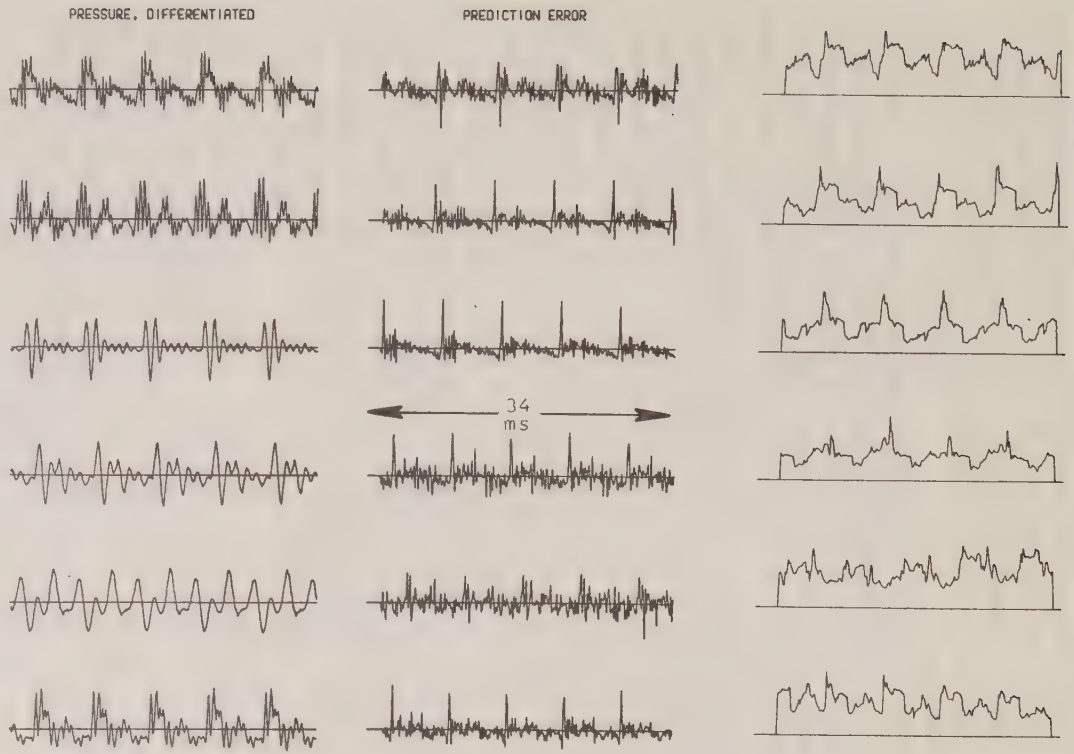


Fig. 10: Differenzierter Schalldruck (links) und Prädiktionsfehler (Mitte) für Vokale /i, e, a*, o, u, y/. Prädiktorgrad p = 12, Hamming-gewichtetes 30-ms-Segment. Rechts: Normierter rms-Prädiktions-fehler auf 2,38-ms-Intervall als Funktion der Intervall-Lage

*/a/ bedeutet immer /ɑ/

finiert sind. Es treten mehrfache Spitzen verschiedener Polarität auf, manchmal (z.B. bei /u/) heben sich die Spitzen gar nicht klar heraus, außerdem muß die Polarität des Signals bekannt sein. Markel [38, 39] benutzt zur Periodenbestimmung nur die Autokorrelationsfunktion dieses Prädiktionsfehlers und verliert damit jede Information über die absolute Lage der Perioden.

Berechnet man die Prädiktoren aus den Autokovarianzmatrizen für sich überlappende Segmente, die kurz gegen eine Periode sind (z.B. 3 ms), so kann man annehmen: In einem Segment, das den Glottisverschlußpunkt enthält, ist der mittlere quadratische Prädiktionsfehler, normiert auf das mittlere Signalquadrat, maximal. In den folgenden Segmenten fällt er stark ab; ein dort berechneter Prädiktor wird relativ gut sein und der zugehörige Prädiktionsfehler scharfe Spitzen beim Glottisverschlußpunkt aufweisen (ähnlich Fig. 14, S. 30). Eine hierauf beruhende automatische Strategie schien nach geeigneter Wahl gewisser Schwellenwerte zunächst sicher zu arbeiten, versagte jedoch bei neuen Sprachaufnahmen. Es traten Mehrfachspitzen des Prädiktionsfehlers auf, oder der rms-Prädiktionsfehler verhielt sich unklar. Eine genaue Untersuchung des rms-Fehlers in Abhängigkeit von der Lage des Segments (Fig. 10, rechts) zeigt zwar meist sprunghaften Anstieg (manchmal mit einer hohen Spitze) und Abfall bei Ein- bzw. Austritt des vermuteten Glottisverschlußpunktes in das bzw. aus dem Analyseintervall, das Gesamtverhalten ist aber doch oft sehr unregelmäßig, besonders bei /u/, und schlecht automatisch auswertbar.

Mehrere rechnerisch äußerst schnelle adaptive Prädiktionsverfahren gibt Maksym an [40]; aus dem Prädiktionsfehler wird durch Pulsformung eine Folge von "Anregungsimpulsen" abgeleitet. Eigene Versuche ergaben jedoch einen relativ großen Prädiktionsfehler, aus dem sich die Pulse keineswegs klarer als für die beschriebenen Verfahren heraushoben (bei /u/ überhaupt nicht); auch der Zusammenhang der Lage der Pulse mit dem Glottisverschluß ist theoretisch wohl noch unsicherer.

Einen anderen Weg weist Sobakin [41]. Fehlerfreie Prädiktion bedeutet lineare Abhängigkeit gewisser Signalsegmente:

Seien s_k Abtastwerte des Signals, $a_1,\ldots,a_p$ die Prädiktorkoeffizienten. Gilt für $k = K_1,\ldots,K_2$

$$s_k = -\sum_{n=1}^{p} a_n s_{k-n} \ ,$$

so sind die Vektoren $\underline{s}_n = (s_{K_1-n},\ s_{K_1+1-n},\ldots,\ s_{K_2-n})^T$, (4:3)
$n = 0,\ldots,p$, linear abhängig:

$$\underline{s}_0 = -\sum_{n=1}^{p} a_n \underline{s}_n \ .$$

Genau dann verschwindet aber die (nie negative) Gram-Determinante $\det(\underline{s}_n^T\ \underline{s}_m)$. Wegen

$$\underline{s}_n^T\ \underline{s}_m = \sum_{k=K_1}^{K_2} s_{k-n} s_{k-m}$$

ist dies die Determinante der $(p+1)\times(p+1)$-Autokovarianzmatrix des Signalsegments von K_1 bis K_2. Sobakin schließt daraus: Wenn die Glottis während des ganzen Segments geschlossen ist, wird die Determinante klein sein, da die Prädiktion und damit die lineare Abhängigkeit besonders gut ist. Ferner sollte der Wert der Determinante als Funktion der zeitlichen Lage des Segments, $\mu(t)$, qualitativ mit den Glottispulsen übereinstimmen. Sobakins Fig. 3c läßt hieran allerdings Zweifel aufkommen: Die Minima der Determinante liegen zu weit rechts bezüglich der vermutlichen Lage der Verschlußintervalle. Eine genauere Untersuchung ist offenbar erforderlich. Drei Einwände erheben sich gegen die ursprüngliche Auffassung: (1) Auch bei nichtverschwindendem Eingangssignal kann die Prädiktion gut sein, solange sich das betrachtete Signalsegment als freier Ausschwingvorgang eines Nur-Pole-Filters auffassen läßt; die aus den Wurzeln des Prädiktorpolynoms berechneten Formanten werden dann jedoch mehr oder weniger fehlerhaft. (2) Es muß genauer geklärt werden, welcher Punkt des Segments am besten als Zeitbezugspunkt zur Definition der Funktion $\mu(t)$ geeignet ist; in [41] wird der Mittelpunkt genommen. (3) Der Wert der Determinante ist nicht nur ein Maß für die lineare Abhängigkeit, sondern wird stark von der Signalamplitude beeinflußt, ist nämlich proportional zu deren $2(p+1)$ten Potenz.

Betrachten wir ein periodisch erregtes Filter, dessen Eingangssignal nur in einem kleinen Teil der Periode $\neq 0$ sei und dessen Ausgangssignal nur mit von Null verschiedenem Fehler linear prädizierbar sein möge. Das Analyseintervall $[K_1,K_2]$

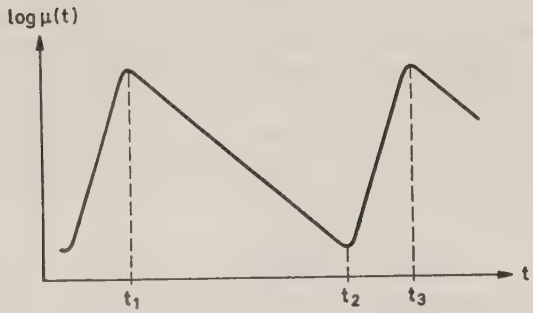


Fig. 11: Schematischer Verlauf des Logarithmus der Determinante der Autokovarianzmatrix

werde in der Zeit verschoben, seine Länge sei konstant und kurz gegen die Signalperiode. Die Ordnung der Prädiktion, p, sei fest. Der Fehler sei in Intervallen, in denen das Eingangssignal Null ist, im (quadratischen) Mittel zur Signalamplitude proportional, die Determinante μ zu einer Potenz des Fehlers. Nehmen wir das Signal als etwa exponentiell abklingend an, so ist der Logarithmus der Determinante eine linear fallende Funktion der Lage des Intervalls (Fig. 11, t_1 bis t_2). Schiebt man das Intervall aber über eine Anregungsstelle hinweg, so wird die Determinante stark wachsen, da die Prädiktion schlecht und außerdem durch die Neuanregung die mittlere Amplitude größer wird (t_2 bis t_3). Das Maximum ist erreicht, wenn schließlich die Anregungsstelle mit dem Intervallanfang zusammenfällt. Bei weiterer Verschiebung nimmt die Determinante wieder ab (nach t_3), da die Prädiktion sich verbessert. Insgesamt sollte der Verlauf von log $\mu(t)$ also etwa sägezahnförmig aussehen. $t_3 - t_2$ ist die Breite des verschobenen Analyseintervalls plus Breite des kleinen Intervalls, in dem die Anregung geschieht. Wählt man also den Anfang des verschobenen Intervalls als Bezugspunkt für dessen Lage, so gibt das Maximum von $\mu(t)$ das Ende der Anregung an.

Trotz der Einfachheit der hier zugrunde gelegten Annahmen zeigt der Verlauf von log $\mu(t)$ bei Sprache i.a. große Ähnlichkeit mit Fig. 11, siehe Fig. 12. Der Schalldruck wurde vor der Aufnahme differenziert (4.2). Dies hebt den Glottisverschlußpunkt stark hervor, so daß die Annahme eines kurzen Anregungsintervalls weniger willkürlich ist, als es zunächst scheint. Der Verlauf hängt nicht sehr vom Rang der Kovarianz-

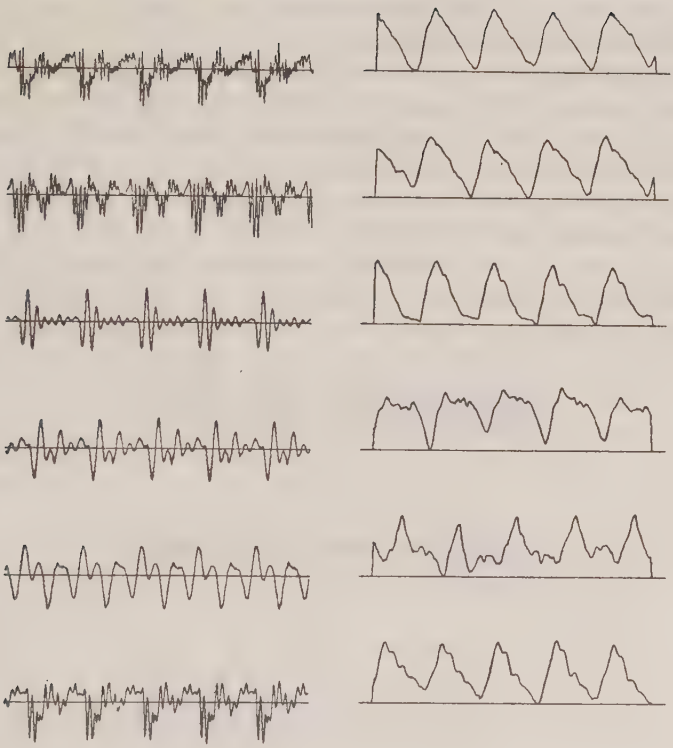


Fig. 12: Logarithmus der Determinante der Autokovarianzmatrix als Funktion der Lage des Analyseintervall-Anfangs (rechts; links differenzierter Schalldruck), p = 12, Intervall-Länge 16 Abtastwerte = 1,81 ms. Vokalaufnahme "TEST DIFF, 2", /i, e, a, o, u, y/

matrix und von der Länge des verschobenen Intervalls ab, die Kurve wird jedoch glatter, wenn diese Größen wachsen. Für p = 12 variiert $\mu(t)$ über bis zu 13 Zehnerpotenzen, bei linearer Auftragung erhielte man also ziemlich schmale Pulse. Das verschobene Intervall hat eine Breite von 15 bis 25 Abtastwerten (1,7 bis 2,9 ms). Auch für sonst schwierige Fälle wie /u/ ergeben sich deutliche Hauptmaxima, wenn auch der Gesamtwertebereich geringer ist.

Die Annahme, daß die Maxima der $\mu(t)$-Kurven das Ende der Anregung und damit den Zeitpunkt des Glottisverschlusses kennzeichnen, kann ohne synchrone Beobachtung der Stimmbandschwingungen mit optischen oder elektrischen Verfahren nicht wirklich bewiesen werden. Jedoch zeigt ein Vergleich mit den oben angedeuteten Prädiktionsfehler-Verfahren, daß die Ergebnisse übereinstimmen, wann immer sie eindeutig sind. Die Spitzen in Fig. 10 Mitte und die sprunghafte rms-Fehler-Verminderung in Fig. 10 rechts liegen an den gleichen Stellen wie die $\mu(t)$-Maxima (Fig. 13). Für /u/ liegen die μ-Maxima dagegen an Stellen, die nach Fig. 10 rechts nicht klar erkennbar sind, aber

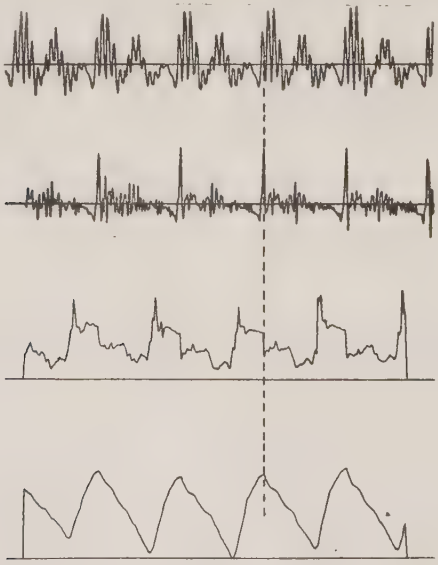


Fig. 13: Vergleich der Kurven nach Fig. 10 und log $\mu(t)$ für /e/. Analyseintervall für die beiden unteren Kurven 2,38 ms = 21 Abtastwerte; Abszisse = Intervallanfang. p = 12

nach Fig. 10 Mitte und nach dem visuellen Eindruck vom differenzierten Schalldruck selbst durchaus als Hauptanregungsstellen in Frage kommen. Eine weitere Möglichkeit, meine Hypothese über die Bedeutung der μ-Maxima zu testen, ist die Inversfilterung der Sprache mit dem Prädiktorpolynom als Übertragungsfunktion, d.h. Berechnung des Prädiktionsfehlers, und Vergleich der erhaltenen "Glottispulse" mit der "theoretischen" Form. Diese Form ist nicht wirklich bekannt, sollte aber eine Anstiegs-, Abfall- und Verschlußphase wie die synthetischen Pulse in Fig. 8 aufweisen. Die Pulsform sollte schlecht herauskommen, wenn der Prädiktor für ein Segment bestimmt wurde, in dem die Glottis sich gerade schloß; bei Verschiebung des Segments in das Verschlußintervall sollte die Pulsform plötzlich ihr bestes Aussehen annehmen. Der Prädiktionsfehler entspricht hier der zweiten Ableitung des glottalen Schallflusses: Für einen hartwandigen Stimmkanal hat die Schallfluß-Übertragungsfunktion nur Pole (von einer stark gedämpften Nullstelle aufgrund des Realteils der Strahlungsadmittanz abgesehen, deren Einfluß bei tiefen Frequenzen vernachlässigbar ist), ihr reziproker Wert ist näherungsweise das Prädiktorpolynom; der Schalldruck im Fernfeld ist zur Ableitung des Schallflusses in der Mundöffnung proportional, ferner wird bei der Aufnahme nochmals differenziert. Bei mitschwingender Rohrwand tritt jedoch mindestens ein weiteres Pol-Nullstellen-Paar in der Übertragungsfunktion auf. Ob das Ergebnis trotzdem sinnvoll ist, wurde mit Hilfe der nach 4.1 synthetisierten Vokale untersucht. Fig. 14 links zeigt das Ergebnis der Inversfilterung mit dem im Intervall mit u(t) = 0 (4:1) gewonnenen Prädiktor und der folgenden zweimaligen Integration, das offensichtlich recht gut ist. Fig. 14 rechts gibt dasselbe für wirkliche Sprache wieder. Überlagerte Reste von Netzbrummen treten infolge der Integration stark hervor (Tiefenanhebung). Nicht immer ist das Verschlußintervall klar hervorgehoben, an seinem Beginn kommen Unregelmäßigkeiten vor, oder es geht ohne Absatz in den nächsten Puls über. Die gezeigten Kurven sind jedoch die jeweils besten, die man bei der Verschiebung des Segments, auf dem der Prädiktor bestimmt wird, erhalten kann; schon eine Verschiebung um wenige Abtastwerte läßt die Ähnlichkeit mit der zu erwartenden Form stark abnehmen. Das Segment wurde im Dialogverfahren unter Benutzung eines Sichtge-

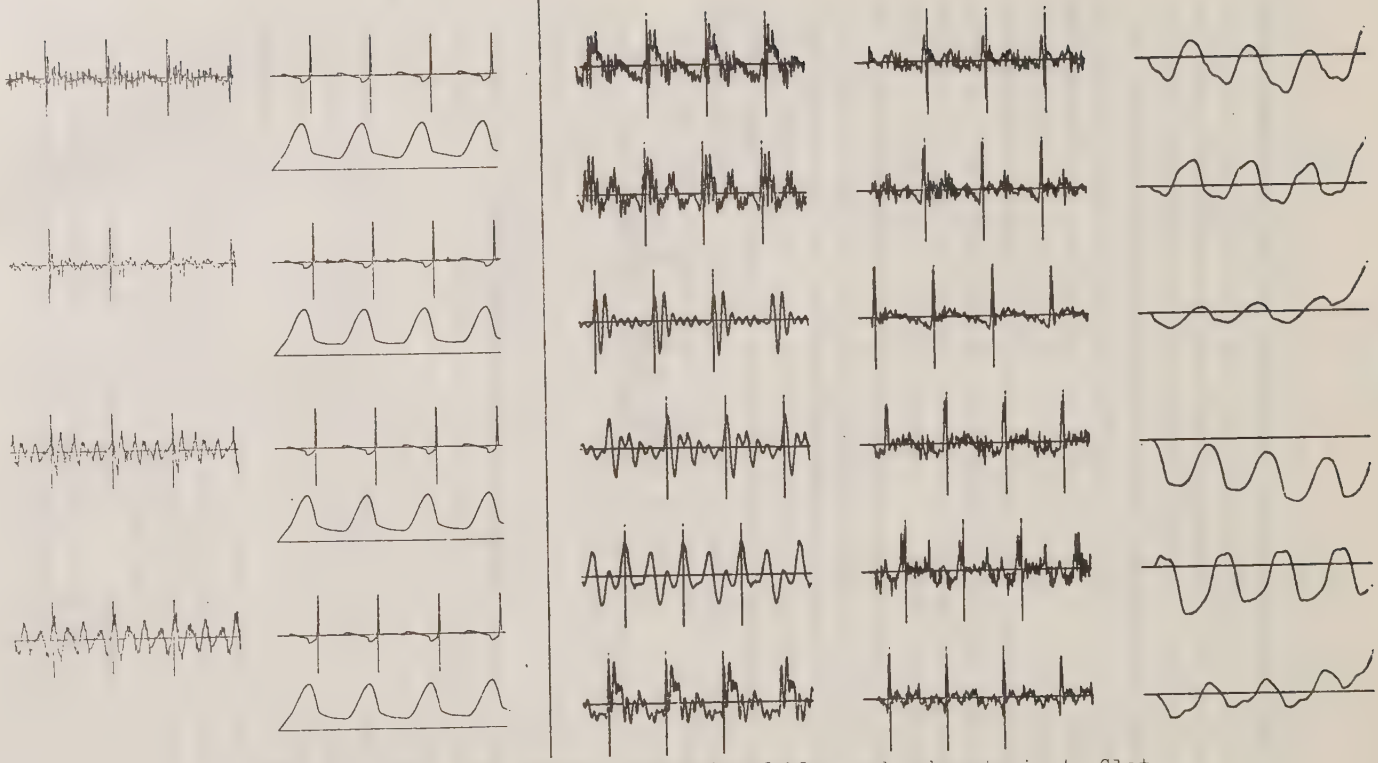


Fig. 14: Differenzierter Schalldruck, Prädiktionsfehler und rekonstruierte Glottiswelle; links synthetische, rechts natürliche Vokale. Der Prädiktor wurde im Intervall nach dem Verschlußpunkt (senkrechte Striche)bestimmt

rätes verschoben; das beste Ergebnis stimmt recht genau mit dem des Determinantenverfahrens überein. - Eine zusätzliche Kontrolle bieten die aus den Wurzeln des Prädiktorpolynoms bestimmten Formanten, die bei Verschiebung des Segments unterschiedlich gut herauskommen.

Um Sobakins Auffassung der Determinante als Maß der linearen Abhängigkeit weiter zu verfolgen, wurde versucht, den Einfluß der Signalamplitude durch verschiedene Normierungen auszuschalten. Das Nächstliegende ist die Ersetzung der Vektoren $\underline{s}_n$ (4:3) durch Einheitsvektoren gleicher Richtung, also der Übergang zur Autokorrelationsmatrix, oder gleichbedeutend damit, die Division der Determinante durch $\prod_{n=0}^{p} |\underline{s}_n|^2$. Das Ergebnis zeigt einen unregelmäßigen Verlauf ohne hinreichend deutliche Hervorhebung des Anregungspunktes. Andere Normierungsversuche waren Division durch $(\sum_{n=0}^{p} |\underline{s}_n|^2)^{p+1}$, $|\underline{s}_0|^{2(p+1)}$, $|\underline{s}_0|^{2p}$, $(\sum_{i=K_1-p}^{K_1-1} s_i^2)^p$, $\left(\sum_{m,n} (\underline{s}_m^T \underline{s}_n)^2\right)^{\frac{p+1}{2}}$. In keinem Fall ergab sich ein Verlauf, der so glatt und leicht automatisch auswertbar war wie die unnormierte Determinante. Aufgrund des obigen Einwandes (1) ist auch nicht unbedingt zu erwarten, daß ein bloßes Maß der linearen Abhängigkeit einen klaren Zusammenhang zum Eingangssignal zeigt. - Übrigens verläuft $\mu(t)$ bei synthetischen Vokalen viel unregelmäßiger als bei natürlichen, da das synthetische Signal Stellen enthält, wo die Determinante wirklich nahezu Null wird.

Ein Nachteil des Verfahrens ist der relativ große Aufwand an Rechenzeit, wenn man vollautomatisch vorgehen will, da für ein mindestens eine Periode umfassendes Signalstück die Funktion μ berechnet, also viele Kovarianzmatrizen und ihre Determinanten gebildet werden müssen. Zum Glück braucht man für jede neue Matrix nur eine Zeile neu zu berechnen, und die Determinante erhält man relativ einfach ohne Zerstörung der Matrix nach dem Choleski-Verfahren [42] (Anhang B). Will man einen zweiten Verschlußpunkt bestimmen, so muß man das Maximum in einem auf den ersten folgenden Intervall suchen, das mindestens eine aber weniger als zwei Perioden umfaßt (Eingabe eines grob geschätzten Periodenwertes ist also nötig). Den n-ten Verschlußpunkt sucht man in einem kleineren Intervall ungefähr eine Periode nach dem (n-1)ten. Das Verfahren wurde

nur an stationären männlichen Vokalen ausprobiert; unstetige Periodenänderungen um den Faktor 2 lassen sich so offenbar nicht erfassen. Nach den zugrunde gelegten Annahmen kommen im übrigen nur kräftige Stimmbandschwingungen mit scharfem Glottisverschluß ("Bruststimme") in Frage. Allgemein kann wohl der Sinn des Verfahrens nicht in der Erzeugung einer Anregungsimpulsfolge für die Übertragung und Synthese längerer Sprachabschnitte liegen, sondern in der Glottisverschlußbestimmung für spezielle Untersuchungen an einzelnen oder wenigen Perioden.

5. Akustische Bestimmung der Querschnittsfunktion

5.1. Bisherige Verfahren

5.1.1. Verwendung der Eingangsimpedanz oder zweier Sätze von Eigenwerten

Die Bestimmung der Querschnittsfunktion aus gewissen akustischen Meßgrößen ist ein Problem der Leitungssynthese. Hierfür geeignete Größen sind z.B. Eingangsimpedanzen am Glottis- und Mundende, Übertragungsfunktionen oder deren Pole und Nullstellen. Nimmt man das Rohr hartwandig und verlustlos an, so bestimmen nach [43, 44, 45] erst zwei Sätze von Eigenwerten der Wellengleichung (2:5) für verschiedene Randbedingungen oder auch die Pole und Nullstellen der Eingangsimpedanz bei bekanntem Abschluß oder die Pole und deren Residuen die Querschnittsfunktion $S(x)$ eindeutig, evtl. bis auf einen Faktor; die Lösungsverfahren benutzen Integralgleichungen. Die Kenntnis der Formantfrequenzen allein genügt also nicht. Nach [46] reicht bereits ein endlicher Abschnitt $[0,T]$ der inversen

Fouriertransformierten $Z_1(t)$ der Eingangsimpedanz, d.h. der Druckantwort auf einen Schallflußimpuls, um einen endlichen Abschnitt [0, cT/2] von S(x) zu bestimmen; die Beziehung ist bijektiv und leuchtet auch anschaulich ein, da die bis zur Zeit T reflektierten Wellen bis zum Punkt x = cT/2 in das Rohr eingedrungen sind. Die Lösung erhält man aus einem Integralgleichungssystem. Einfachere Lösungsverfahren ergeben sich für diskrete Approximationen der Leitung. Bei Stufenleitungen läßt sich die Querschnittsfunktion S_n direkt aus den Ausbreitungsgleichungen (2:14) rekonstruieren [47]: Seien alle u_k^n, $p_k^n = 0$ für $k < 0$. Nach (2:14) für $k = -1$ ist dann $S_n = u_0^n/p_0^n$, was bei Kenntnis von u_k^n, p_k^n, $k = 0,\ldots,M$ die Berechnung von S_n und nach (2:14) auch von u_k^{n+1}, p_k^{n+1}, $k = 0,\ldots,M-1$ gestattet. Sind also u_k^1, p_k^1, $k = 0,\ldots,N$ bekannt, so lassen sich die Querschnitte $S_1,\ldots,S_{N+1}$ nacheinander ermitteln. Auch hier wird somit wieder aus einem Zeitabschnitt T des Eingangssignals auf die Querschnittsfunktion einer Rohrlänge cT/2 geschlossen. Ein ähnliches Verfahren geben Paige und Zue [48] an; es ist völlig äquivalent zu dem gerade erwähnten von Atal, aber im Frequenzbereich formuliert und daher rechnerisch schwieriger zu handhaben. Das Entsprechende läßt sich auch für eine diskrete LC-Leiter durchführen [49, 50], wobei die Konstanten in der Kettenbruchdarstellung der Impedanz bestimmt werden.

Anders gehen Schroeder und Mermelstein [51, 52, 53] vor: Sind f_n, g_n zwei Sätze von Resonanzfrequenzen für weichen bzw. harten Rohrabschluß (nach (2:2) identisch mit Nullstellen bzw. Polen der Eingangsimpedanz des Mundes bei harter Glottis) und entwickelt man den Logarithmus der Querschnittsfunktion in eine Kosinusreihe

$$\log S(x) = \sum_{n=0}^{\infty} a_n \cos \frac{n\pi x}{L}, \quad L \text{ Rohrlänge}, \tag{5:1}$$

so gilt für kleine Abweichungen vom homogenen Rohr ($|a_n| \ll 1$, $n \neq 0$):

$$\frac{\delta f_n}{f_n} = -\frac{a_{2n-1}}{2}, \qquad \frac{\delta g_n}{g_n} = -\frac{a_{2n}}{2}. \tag{5:2}$$

Bei großen Abweichungen von der Homogenität liegt kein linearer Zusammenhang mehr vor, jedoch lassen sich die a_n iterativ durch Anpassung der f_n, g_n oder der monoton wachsenden Phase der Eingangsadmittanz [54] an die Meßwerte bestimmen. a_0 bleibt unbestimmt, wenn nicht auch der Betrag der Impedanz benutzt wird.

All diese Verfahren werfen das grundsätzliche Problem auf, daß die Eingangsimpedanz für alle Frequenzen bekannt sein muß; physikalisch sinnvoll sind aber nur Frequenzen bis etwa 4 kHz, da oberhalb keine eindimensionale Ausbreitung mehr vorliegt. Beim Schroeder-Mermelstein-Verfahren sieht man, daß bei Benutzung nur endlich vieler Resonanzen ($n = 1,\ldots,N$) im linearen Fall (5:2) genau die ersten a_n, $n = 1,\ldots,2N$, bestimmbar sind. Für den allgemeinen Fall ist das nicht bewiesen, sieht aber empirisch sinnvoll aus. - Bei den nichtiterativen Verfahren werden nötigenfalls für die fehlenden höheren Resonanzen künstlich diejenigen eines "kanonischen" (z.B. homogenen) Rohres eingesetzt, man erhält dann eine geglättete Version der Querschnittsfunktion. Bei diskret approximierten Leitungen kann man auch von vornherein die Segmentierung hinreichend grob wählen; die Zahl der rekonstruierten Segmente ist gleich der Gesamtzahl der Pole und Nullstellen. Allgemein beträgt das Auflösungsvermögen bei Bandbegrenzung auf 4 kHz bestenfalls 2 cm.

Zwei weitere Schwierigkeiten sind, daß bei allen Verfahren, die im Frequenzbereich arbeiten oder die Querschnittsfunktion durch ein Modell mit zu bestimmenden Parametern nach Art von (5:1) darstellen, die Rohrlänge und die Abschlußimpedanz am anderen Rohrende (zumindest ihr Phasengang) bekannt sein müssen. Beim Arbeiten im Zeitbereich [45, 47] ergeben sich diese Probleme dagegen nicht; es sei denn, höherfrequente Resonanzen müssen künstlich hinzugefügt werden. Die Länge läßt sich grob aus der Folge der Pole und Nullstellen abschätzen [51 bis 54], diese nähern sich asymptotisch denen eines homogenen Rohres gleicher Länge. Nach [55] gilt auch empirisch: Bei richtiger Länge ist die Varianz von $\log S(x)$ minimal, wenn $S(x)$ die errechnete Querschnittsfunktion ist.

Das einzige Verfahren, das bisher auf einen wirklichen Stimmkanal angewandt wurde, scheint das von Schroeder, Mermelstein und West zu sein. Da sich die Eingangsimpedanz am Glottisende nicht direkt messen läßt, mußte die des Mundes benutzt werden, wozu die Versuchsperson den Mund an ein Kundtsches Rohr mit Mundstück hielt. Der komplexe Reflexionsfaktor des Mundes wurde aus Impulsechos mit einem akustischen Richtkoppler ermittelt. Wirkliches Sprechen mit Phonation war nicht möglich, Glottis und Velum mußten geschlossen sein, der Mund

konnte nicht weit geöffnet werden. Immerhin ließ sich eine gewisse Hörrückkopplung durch Synthese eines zur gemessenen Impedanz passenden "Sprachsignals" erreichen.

5.1.2. Verwendung des Sprachsignals

Die praktischen Schwierigkeiten der Impedanzmeßverfahren legen es nahe, mit dem Sprachsignal, genauer der Übertragungsfunktion H_u, selbst zu arbeiten. Wie schon gesagt, reicht die Kenntnis der Formantfrequenzen nicht aus. Dennoch enthält das Sprachsignal (bei hinreichender Kenntnis des Glottissignals) im Prinzip genügend viel Information, um die Querschnittsfunktion (evtl. bis auf einen Faktor) zu bestimmen, da bei vorhandener Dämpfung, z.B. durch die Strahlungslast, die Formantbandbreiten die fehlende Information enthalten. Kenntnis der Übertragungsfunktion erlaubt es bei verlustfreiem Rohr, die Eingangsimpedanzen am Mund- und Glottisende, Z_2 bzw. Z_1, zu berechnen, wenn die Strahlungsadmittanz Y_r als RL-Parallelschaltung mit konstanten R und L approximiert wird. Z_2 erhält man nach (2:2) durch Zerlegung des Nenners von H_u unter Beachtung, daß $A_{22}(s)$ nur gerade, $A_{21}(s)$ nur ungerade Potenzen von s oder ω enthält [50] bzw. daß nach (2:24) $\hat{A}^x_{22}(z) = \hat{A}_{22}(z)$, $\hat{A}^x_{21}(z) = -\hat{A}_{21}(z)$ ist (das $\hat{}$ kennzeichnet die Weglassung des Laufzeitfaktors $z^{N/2}$). Z_1 folgt dagegen aus der Leistungsbilanz: Ist G_r der Realteil der Strahlungsadmittanz, so wird

$$|P_2|^2 G_r = |U_1|^2 \mathrm{Re}(Z_1); \qquad (5:3)$$

bei nicht zu hohen Frequenzen ($\omega \ll 1/G_r L_r$) ist der Druck in der Mundöffnung, P_2, etwa proportional dem Druck im Strahlungsfeld, $P_2/U_1 = Y_r^{-1} H_u \approx sL_r H_u$ also die Gesamtübertragungsfunktion. Dies liefert $\mathrm{Re}(Z_1)$; $\mathrm{Im}(Z_1)$ ist aus Kausalitätsgründen die Hilberttransformierte von $\mathrm{Re}(Z_1)$ (Kramers-Kronig-Relationen)[56]. Geht man in den Zeitbereich über [47], so wird $|P_2|^2/|U_1|^2$ zur Autokorrelationsfunktion R(t) der Impulsantwort des Ausgangsdruckes auf den Eingangsschallfluß, $\mathrm{Re}(Z_1)$ zum geraden Anteil von $Z_1(t)$, daher gilt:

$$Z_1(t) = \begin{cases} 0 & , t < 0 \\ G_r R(0) & , t = 0 \\ 2G_r R(t), & t > 0 \end{cases} . \qquad (5:4)$$

Es besteht also ein umkehrbar eindeutiger Zusammenhang zwi-

schen $R(t) \longleftrightarrow Z_1(t) \longleftrightarrow S(ct/2), \quad 0 \leq t \leq T,$
bzw. diskret $R_n \longleftrightarrow Z_{1n} \longleftrightarrow S_{n+1}, \quad n = 0,\ldots,N.$ (5:5)

Die Rohrlänge braucht nicht bekannt zu sein. Da $G_r \propto$ Querschnitt des Mundes, ist S(x) nur bis auf einen Faktor bestimmbar, selbst wenn R(t) absolut bekannt wäre. Kenntnis der Autokorrelationsfunktion ist äquivalent zur Kenntnis des Leistungsspektrums bei Deltaimpulsanregung oder eines Prädiktorpolynoms hinreichenden Grades (Anhang A), auch hieraus läßt sich also die Querschnittsfunktion ermitteln.

Für eine Stufenleitung kann man die Querschnittsfunktion auch direkt aus einem Stück der Autokorrelationsfunktion oder dem Prädiktorpolynom berechnen, ohne daß man die Eingangsimpedanz als Zwischenergebnis benötigt (Atal [26, 47], Wakita [27]). Diese Verfahren sollen im nächsten Kapitel kritisch untersucht werden.

Die für die Impedanzverfahren genannten Probleme gelten entsprechend auch für die Sprachsignal-Verfahren. Hinzu kommen die Schwierigkeiten, daß das glottale Eingangssignal nur grob bekannt ist und daß sich die Bandbreiten, die ja die Hälfte der benötigten Information enthalten, schwer genau bestimmen lassen. Denn das Spektrum besteht näherungsweise aus diskreten Linien, deren Abstand größer als die Bandbreiten sein kann; Verschmierung durch ein endliches Zeitfenster oder absichtliche Glättung liefert jedoch viel zu große Bandbreiten, verglichen mit den direkten Messungen von [28, 31, 57] mit künstlicher Anregung des Stimmkanals. Eigentlich müßte das Faltungsprodukt von Glottiswelle und Stimmkanal-Impulsantwort wieder dekonvolviert werden, was aber nur mit viel Information über die Glottiswelle möglich wäre. Eine weitverbreitete grobe Näherung für den Gesamtfrequenzgang ist ein Abfall mit etwa 6 dB/Okt.; auch eine Darstellung durch Pole in der s- oder z-Ebene kommt vor [26]. Cepstrale oder komplex-cepstrale Glättung kann dann die hochquefrenten Strukturen beseitigen[24, 58, 59], verschmiert aber auch die Formanten. Besser scheint hier die Verwendung von prädiktiven Verfahren [26, 38, 39, 60] zur Bestimmung der Übertragungsfunktion (Anhang A), die Bandbreiten sind aber auch dabei mit erheblicher Unsicherheit behaftet. Denn die periodische Änderung der Glottisimpedanz bewirkt eine periodische Änderung der Formantfrequenzen und -bandbreiten [15] (vgl. 2.1), also ebenfalls eine Verschmierung; wenn das Zeit-

fenster mehr als einen Bruchteil einer Grundton-Periode enthält. Hinzu kommt, daß der Einfluß der querschnitts- und frequenzabhängigen Rohrdämpfung gegenüber der Strahlungsdämpfung schwer zu berücksichtigen ist; hier ist man auf empirische Faustformeln angewiesen. Fujisaki, Obata und Tazaki [50] z.B. messen die Bandbreiten überhaupt nicht, sondern benutzen aus den Werten einer homogenen Modell-Leitung interpolierte Werte, was aber doch recht willkürlich erscheint. Auch die empirische Formel für den Zusammenhang zwischen Formantfrequenzen und -bandbreiten nach Fant [32] wäre vielleicht brauchbar. Im Grunde handelt es sich darum, die bei bloßer Kenntnis der Formantfrequenzen noch unendliche Lösungsmannigfaltigkeit empirisch auf eine Lösung zu reduzieren. Statt die Bandbreiten zu schätzen, könnte man auch direkt die Menge der möglichen Querschnittsfunktionen einschränken. Mermelstein und Schroeder machen z.B. davon Gebrauch, daß die Querschnittsfunktionen oft (außer bei /u/) ungefähr die Relation

$$S(L-x) = 1/S(x), \quad L \text{ Rohrlänge}, \tag{5:6}$$

erfüllen und setzen in der Kosinusreihe (5:1) alle a_n mit geradem n gleich Null. Pinson [61] iteriert die Querschnitte S_n unter Anpassung allein der Formantfrequenzen an; Eindeutigkeit wird erreicht, indem bei jedem Iterationsschritt der Korrekturvektor $\Delta\underline{S}$ minimaler Länge gewählt wird, eine physikalisch schwer zu begründende ad-hoc-Annahme.

Im nächsten Abschnitt soll untersucht werden, wie stark die aus dem Sprachsignal berechnete Querschnittsfunktion von Annahmen über die physikalischen Ursachen der Dämpfung und vom Frequenzgang abhängt. Der Vergleich Atal-Wakita zeigt diese Abhängigkeiten besonders klar. Das benutzte Rechenverfahren dürfte für diese Betrachtung jedoch keine Rolle spielen; hierfür kann nur die zugrunde liegende physikalische Modellvorstellung entscheidend sein. Von allen genannten Verfahren scheint übrigens nur Wakitas an wirklicher Sprache getestet worden zu sein; zumindest haben andere Autoren m.W. keine Ergebnisse veröffentlicht (Stand: Juni 1974).

5.2. Kritische Untersuchung der Verfahren von Atal und Wakita

5.2.1. Die beiden Verfahren

Im vorigen Kapitel wurden die Verfahren von Atal [26, 47] und Wakita [27] erwähnt, die die unmittelbare Bestimmung der Querschnittsfunktion S_n einer Stufenleitung aus dem Sprachsignal gestatten sollen. Diese seien zunächst kurz dargestellt. Beide gehen davon aus, daß sich in den Gleichungen (A:10) bis (A:13), die einen Abschnitt R_n/R_0, $n = 1,\ldots,N$, der Autokorrelationsfunktion der Impulsantwort mit den Prädiktorpolynomen n-ten Grades, $A_n(z)$, $n = 1,\ldots,N$, verknüpfen, die Koeffizienten a_n^n mit den Reflexionsfaktoren des Rohres nach (2:6) identifizieren lassen, woraus man die Querschnittsfunktion bis auf einen Faktor erhält. Beide benutzen jedoch verschiedene Randbedingungen.

1.) Atals Verfahren. Die Glottis wird schallhart angesetzt, d.h. der Schallfluß eingeprägt. Die Strahlungsadmittanz am Mundende habe einen positiven, konstanten Realteil G_r. Hat das Rohr N Segmente, so ist nach (5:5) der Abschnitt R_k/R_0, $k = 1,\ldots,N-1$ der AKF von $\mathcal{Z}^{-1}(P_2/U_1) = \mathcal{Z}^{-1}(Y_r^{-1}H_u)$ und damit die Prädiktorpolynome A_n, $n \leq N-1$, nur von $S_1,\ldots,S_N$ abhängig, nicht von der Strahlungsadmittanz Y_r (diese Überlegung fehlt in [26], ist aber in [47], wo die Herleitung ganz anders über die Kettenmatrixelemente geschieht, implizit enthalten). Setzen wir diese $= S_N$ (reflexionsfreier Abschluß; in [26] ausschließlich betrachtet), so ist (vgl. (2:2), (2:25)) die Übertragungsfunktion $Y_r^{-1}H_u \propto H_u \propto C_{N-1}^{-1}$ und hat nur Pole (N-1 Stück), so daß nach Anhang A C_{N-1} bis auf einen Faktor das Prädiktorpolynom A_{N-1} ist. Die Beziehungen (2:23) für die inversen Schallfluß-Übertragungsfunktionen C_n stimmen formal mit den Prädiktor-Beziehungen (A:11) überein, wenn man r_n mit a_n^n identifiziert, also $A_n = C_n$, $n = 1,\ldots,N-1$. Kennt man daher A_{N-1} oder R_k/R_0, $k = 1,\ldots,N-1$, so erhält man $r_1,\ldots,r_{N-1}$ aus (A:13) bzw. (A:11), (A:12). Weitere r_n, $n \geq N$, simulieren lediglich das Verhalten der Strahlungsadmittanz.

2.) Wakitas Verfahren. Die Strahlungsadmittanz ist unendlich (Kurzschluß), also der Reflexionsfaktor $r_N = 1$, wenn das Rohr N Segmente hat. Dagegen liegt ein dämpfungserzeugender Leitwert S_0 parallel zur Glottis, was formal einem zusätzlichen Segment S_0 entspricht, in dem die hinlaufende Welle

$U_0^+ = z^{1/2}U_{glottis}/2$ eingeprägt, die rücklaufende U_0^- absorbiert wird. U_{N+1}^+ ist die abgestrahlte Welle, U_{N+1}^- sei Null. Definiert man inverse Übertragungsfunktionen

$$D_n^{\pm} = (U_n^{\pm}/U_{N+1}^+) \prod_{\nu=n}^{N} ((1+r_\nu)z^{-1/2}), \qquad (5:7)$$

so gilt nach (2:17)

$$\begin{pmatrix} D_n^+ \\ D_n^- \end{pmatrix} = Q_n \begin{pmatrix} D_{n+1}^+ \\ D_{n+1}^- \end{pmatrix}, \quad D_N^+ = 1, \; D_N^- = -z^{-1}. \qquad (5:8)$$

Dies sind aber wieder die Prädiktorgleichungen (A:11), wenn man $A_n = D_{N-n}^+$, $A_n^x = -zD_{N-n}^-$, $a_n^n = r_{N-n}$ setzt. $1/A_N = 1/D_0^+$ ist die Schallfluß-Übertragungsfunktion des Gesamtrohres. Wieder bekommt man also bei Kenntnis des Prädiktors A_{N-1} oder von R_k/R_0, $k = 1,\ldots,N-1$, der AKF von $\mathfrak{F}^{-1}(H_u)$, Reflexionsfaktoren $r_1,\ldots,r_{N-1}$; jedoch in umgekehrter Reihenfolge wie bei Atal. A_N liefert das die Dämpfung repräsentierende r_0, alle A_n mit $n > N$ sollten $= A_N$ sein ($r_m = 0$, $m < 0$). Wakita selbst numeriert die Segmente übrigens vom Mund ausgehend, seine Notation ist anders.

5.2.2. Theoretische Überlegungen

Es muß untersucht werden, welche der beiden Annahmen über die Randbedingungen sinnvoller ist. Wakita streitet die Anwendbarkeit von Atals Verfahren ab. Tatsächlich ist der Hauptunterschied beider Verfahren, die ja genau denselben Algorithmus verwenden, folgender: Da die S_n nach (2:6) aus den r_n gebildet werden, erhält man je nach Auffassung der Randbedingungen bzw. der Laufrichtung des Indexes n S_n oder S_{N-n}^{-1}; bei Auftragung von log S_n entspricht das einer Rotation der Kurve um 180°. Ein weiterer Unterschied liegt im Gesamtfrequenzgang des benutzten Signals: Die Impulsantwort, deren AKF verwendet wird, ist bei Atal die Antwort des Ausgangsschalldrucks, bei Wakita die des Ausgangsschallflusses auf einen Schallfluß-Einheitsimpuls am Eingang; bei Atal muß das Spektrum also im Mittel mit 6 dB/Okt. ansteigen, bei Wakita flach sein.

Beide Autoren machen unrealistische Annahmen über die Dämpfungsquellen: Bei Atal existiert nur die Strahlungsdämpfung, bei Wakita nur ein dämpfender Leitwert parallel zum Rohreingang. Atals Auffassung erscheint für offene Vokale sinnvoller; bei engen Einschnürungen in Mundnähe wird jedoch

die Strahlungsdämpfung klein verglichen mit den Viskositäts- und Wärmeleitungsverlusten, bei tiefen Frequenzen spielen auch die Wandverluste erheblich mit. Hier ist vielleicht Wakitas Auffassung besser; dennoch sieht seine Zusammenfassung aller Verluste in einen diskreten Querleitwert am Eingang zweifelhaft aus, da, von den Wandverlusten abgesehen, gerade enge Stellen weiter vorn im Rohr die Hauptdämpfung erzeugen.

Wakita zeigt im Gegensatz zu Atal Ergebnisse für wirkliche Sprachaufnahmen, sogar für Lautübergänge, die vernünftig aussehen. Er geht von der AKF eines Hamming-gewichteten 20-ms-Segments differenzierter Sprache (f_s = 7 kHz) aus. Dabei ergibt sich das weitere Problem, ob diese AKF in ihrem Anfangsstück mit der AKF der Impulsantwort übereinstimmt. Da nach [60] der durch dieses Anfangsstück gegebene Prädiktor gerade die spektralen Maxima, d.h. die Einhüllende, gut approximiert, könnte das der Fall sein, wenn sich das Spektrum des Glottisschallflusses gut durch einen 12-dB/Okt.-Abfall darstellen läßt (6 dB/Okt. werden durch die Abstrahlung, weitere 6 dB/Okt. durch die Differenzierung wieder ausgeglichen). Jedoch werden die Bandbreiten aus den in 5.1.2 genannten Gründen zu groß bzw. die spektralen Spitzen zu klein herauskommen. Die Bandbreiten sind außerdem auch theoretisch falsch: Für ein homogenes Rohr beispielsweise liefert die Annahme von S_0 als einziger Dämpfungsquelle konstante Bandbreiten bzw. Resonanzmaxima; in Wirklichkeit nimmt die Höhe der Maxima jedoch mit zunehmender Formantnummer ab, da Strahlungs- und innere Verluste größer werden. Evtl. ist also eine weitere Höhenanhebung nötig. - Bei Atals Verfahren wäre der Schalldruck nicht nur einmal wie bei Wakita, sondern zweimal zu differenzieren und dann die AKF zu bilden. Da jedoch die oben erwähnte Formantverschmierung (5.1.2) die Güte der niedrigen Formanten und so die Höhe der zugehörigen spektralen Maxima stark verringern dürfte, könnte sich das als zu starke Höhenanhebung erweisen. Die beiden Verfahren sollten sich also im benötigten Frequenzgang nicht so sehr unterscheiden.

Der Zusammenhang zwischen den benutzten Prädiktorpolynomen und der Übertragungsfunktion ist auch im Idealfall durchaus nicht trivial. Lediglich wenn Wakitas Modell exakt wäre, wäre $A_n \propto H_u^{-1}$ für alle $n \geq N$. Schon für das homogene Rohr hat aber das eigentlich nur benötigte A_{N-1} keine Ähnlichkeit mehr

mit H_u^{-1}; es ist nämlich für $n \leq N-1$ $A_n = 1$. Dies gilt auch bei Atals Modell. Hier unterscheidet sich aber sogar A_N erheblich von der inversen Übertragungsfunktion; denn A_N^{-1} ist ja die Übertragungsfunktion, die das Rohr hätte, wenn es mit dem Reflexionsfaktor r_N reell abgeschlossen wäre. Da $P_2/U_1 = Y_r^{-1} H_u$ eine Nullstelle bei $\omega = 0$ hat, kann die Übertragungsfunktion auch gar nicht $\propto A_N^{-1}$ sein. Der Unterschied zwischen A_N^{-1} und P_2/U_1 bzw. H_u soll wieder am homogenen Rohr plausibel gemacht werden. Bei Abschluß mit einem reellen Leitwert S_{N+1} wären alle Formantbandbreiten bzw. die Maxima von H_u gleich; bei Abschluß mit einer wirklichen Strahlungsimpedanz dagegen nehmen mit wachsender Formantnummer die Bandbreiten zu und die Maxima ab. Diese Abnahme wird durch Übergang zu $Y_r^{-1} H_u$ wieder ausgeglichen. Da der aus der AKF gewonnene Prädiktor gerade die spektralen Maxima bevorzugt berücksichtigt, erhält man bei geeignet gewähltem S_{N+1} den gleichen Prädiktor für beide Abschlüsse. Die Wurzeln des Prädiktors A_N liefern also nicht die richtigen Formantwerte; die Frequenzen werden zwar nicht allzu sehr abweichen, wenn $r_N > 0$ ist; die Bandbreiten werden jedoch bei den tiefen Formanten zu groß sein.

5.2.3. Modellrechnungen und Experimente

Da beide Modellvorstellungen mit erheblichen Unsicherheiten bzgl. des Gesamtfrequenzgangs behaftet sind, soll dessen Einfluß kurz untersucht werden, ebenso die Wirkung einer zusätzlichen Abweichung von den Modellannahmen, nämlich der mitschwingenden Wand. Wir betrachten zunächst ein reell abgeschlossenes homogenes Rohr ($r_n = 0{,}94$ für $n = N$, $r_n = 0$ sonst) mit $N = 9$ Segmenten von $d = 2$ cm Länge. Hier kommt von vornherein nur Atals Auffassung in Betracht. Aus der exakten AKF der Impulsantwort wird die Querschnittsfunktion bestimmt. Zweitens wird eine Höhenanhebung von 6 dB/Okt. vorgenommen ($R_k \mapsto 2R_k - R_{k-1} - R_{k+1}$), drittens mit der exakten AKF eines Rohres mit einem RL-Querzweig nach 3.2 am Eingang gearbeitet. Die drei Ergebnisse zeigt Fig. 15, 1 bis 3; bei Fall 2 und 3 bestehen starke Abweichungen von der richtigen Lösung 1. Eine leichte Rechnung nach (A:11), (A:12) zeigt, daß in Kurve 2 $S_n \propto n(n+1)$, $n < N$, statt $S_n = \text{const}$. Die Frequenzgang-Abhängigkeit ist also erheblich, wie schon Wakita feststellte. Der

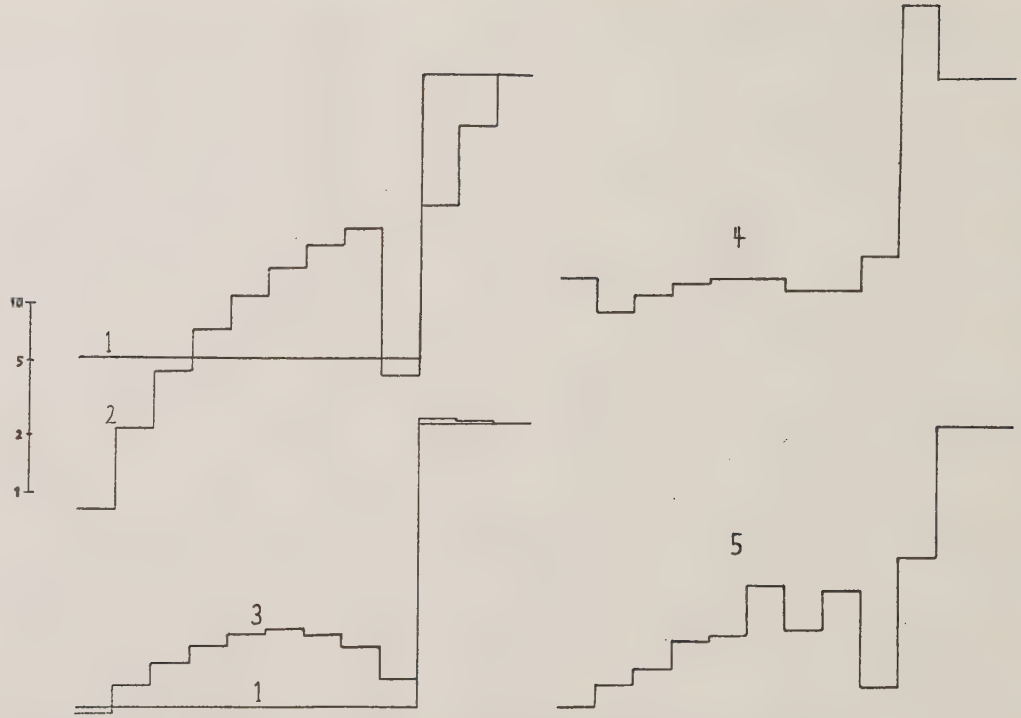


Fig. 15: Bestimmung der Querschnittsfunktion eines homogenen, reell abgeschlossenen Rohres, links (1-3) aus der AKF der Impulsantwort, rechts (4-5) aus Prädiktor 10. Grades nach dem Kovarianzverfahren. Kurve 1 entspricht dem Originalrohr, im übrigen siehe Text

Einfluß des "Wandimpedanz"-Querzweiges besteht in einer Deformation des Rohres; sein Anfangsteil wird etwa exponentiell aufgeweitet (Kurve 3). Diese Wirkung wurde auch am homogenen Rohr und an Fant-Querschnittsfunktionen, die mit einem Strahlungswiderstand (3:3) abgeschlossen waren (Atals Auffassung), untersucht (Fig. 16); N = 17 bis 20, d = 1 cm, die exakte AKF der Druck-Impulsantwort lag zugrunde. Bei offenen Vokalen (F_1 hoch) ist der Fehler erträglich, bei Einschnürung im Vorderteil (F_1 tief) sind die Ergebnisse dagegen völlig falsch. Das ist zu erwarten: Bei tiefem F_1 ist der Wandeinfluß groß; er wird im Ergebnis offenbar durch zusätzliche Strahlungsdämpfung (große Mundöffnung) simuliert. Die Wirkung scheint in einer Verschiebung der Kurve zum Mund hin und einer trichterartigen Aufweitung zu bestehen.

Ergebnisse für nach 4.1 synthetisierte (a) und für natürliche (b) Vokale mit d = 2 cm zeigt Fig. 17. Die Vokale "FANT IEAOUY" sind /i, e, a, o, u, ɨ/ (synthetisch), die Vokale "TEST SPRACHE" und "TEST DIFF, 2" sind /u, o, o, a, e, i, y, ø, ə/ und /i, e, a, o, u, y/, gesprochen vom Autor und aufgenommen nach 4.2. Die ursprüngliche Abtastfrequenz 35,3 kHz wurde zuerst nach Tiefpaßfilterung mit (4:2) auf ein Viertel (8825 Hz) herabgesetzt, entsprechend d = 2 cm. Den Ergebnissen liegt die AKF eines Hamming-gewichteten 29-ms-Abschnitts (256 Abtastwerte) zugrunde. Die Höhenanhebung beträgt 6 und 12 dB/Okt. Die zweimalige Differentiation bewirkt hauptsächlich eine Verschiebung der Querschnittsfunktion um 1 bis 2 Segmente nach vorn, von weiteren Änderungen abgesehen. Ohne Querzweig sollten die 12-dB/Okt.-Proben die (nach Atal) theoretisch besseren sein; jedoch bewirkt der Querzweig (außer bei /a/ und /o/) wieder ein trichterartig aufgeweitetes Anfangsstück und eine Vorwärtsverschiebung. Die 6-dB/Okt.-Proben sehen oft besser aus. Eine Entscheidung zwischen Atals und Wakitas Auffassung fällt schwer, da die Querschnittsfunktionen näherungsweise (5:6) erfüllen, sich also bei Rotation um 180° wenig ändern. Bei offenen synthetischen Vokalen ist Atals Auffassung, wie zu erwarten, besser; bei /u/ dagegen kommt der Mundverschluß nie heraus, was an der gegenüber dem Atalschen Modell viel zu hohen Dämpfung der tiefen Formanten liegen dürfte. Wakitas Auffassung liefert hier den Mundverschluß. Die Güte der Ergebnisse für natürliche Vokale ist schwer zu entscheiden. Die gün-

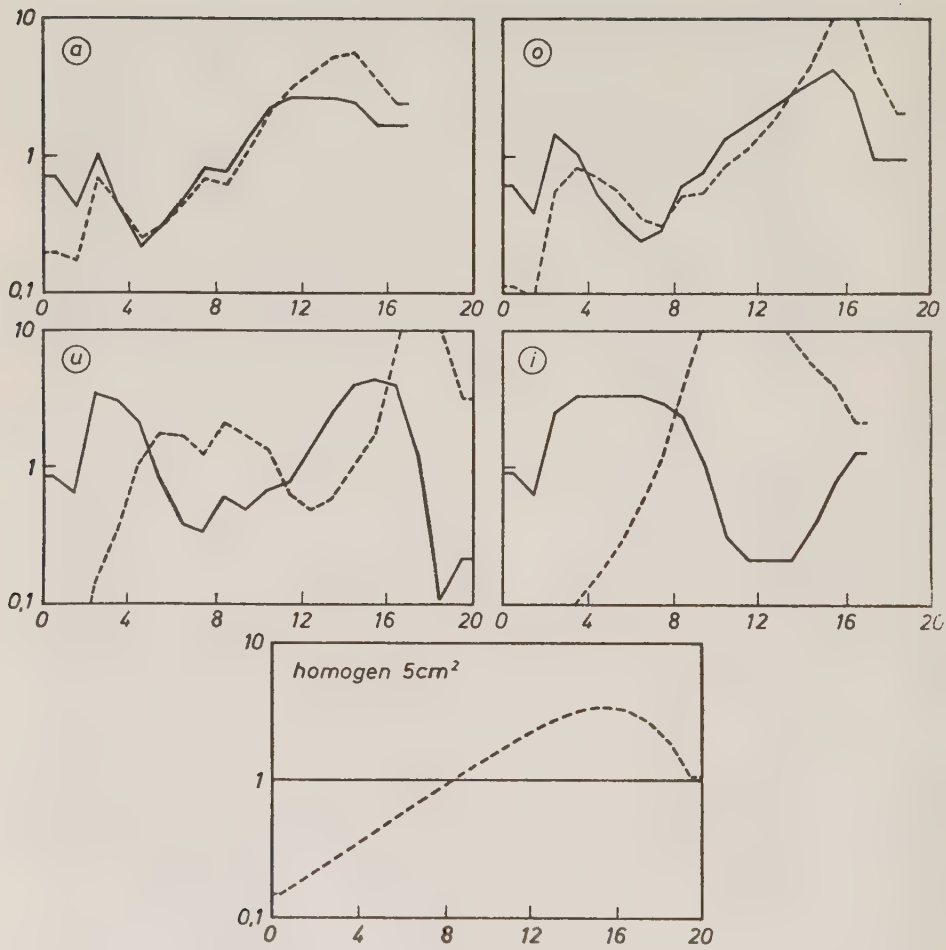


Fig. 16: Aus der exakten AKF der Impulsantwort nach Atal rekonstruierte Querschnittsfunktionen (gestrichelt); Originalrohr (durchgezogen) mit RL-Querzweig (Fig. 6 a) am Eingang und Strahlungsimpedanz am Ausgang. Ordinate jeweils auf Mittelwert 1 normiert. Segmentlänge 1 cm

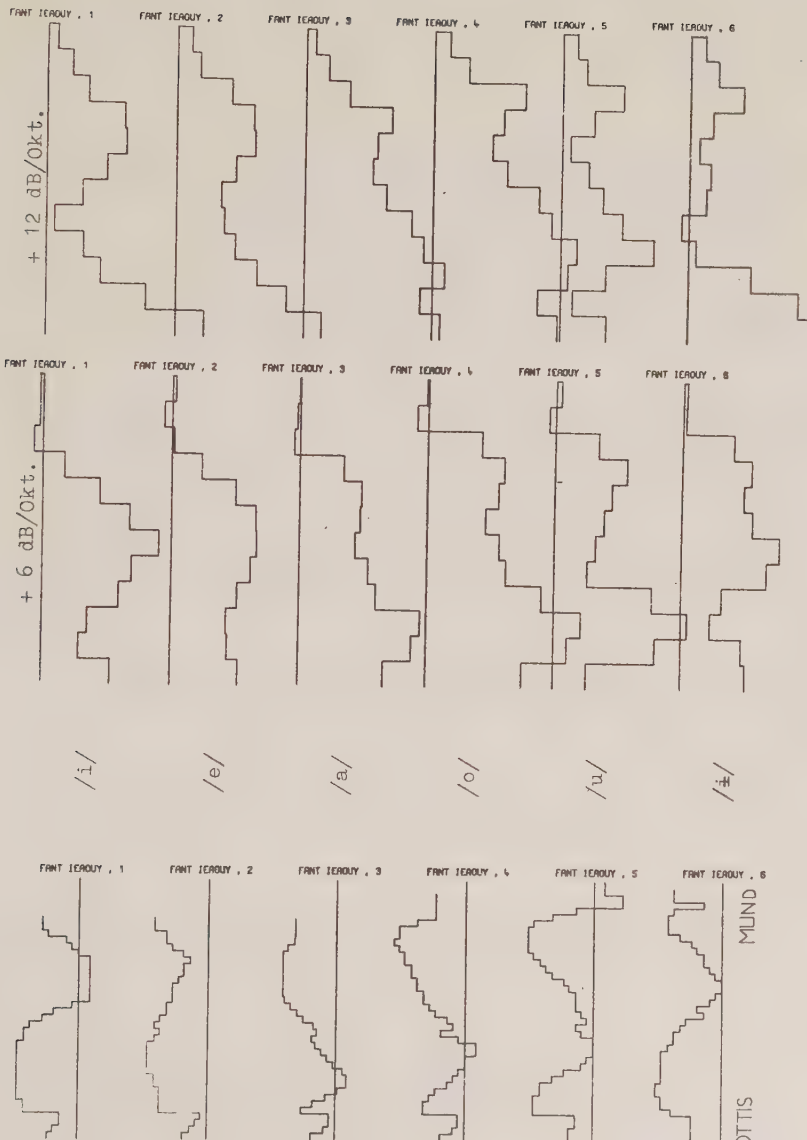


Fig. 17 a: Bestimmung der Querschnittsfunktion aus der AKF eines Hamming-gewichteten 29-ms-Abschnitts. Fant-Vokale (links Originale, d = 0,5 cm), d = 2 cm, Schalldruck ein- und zweimal differenziert (Mitte und rechts). Darstellung nach Atal; nach Wakita Kurven um 180° drehen! Es wurden zwischen

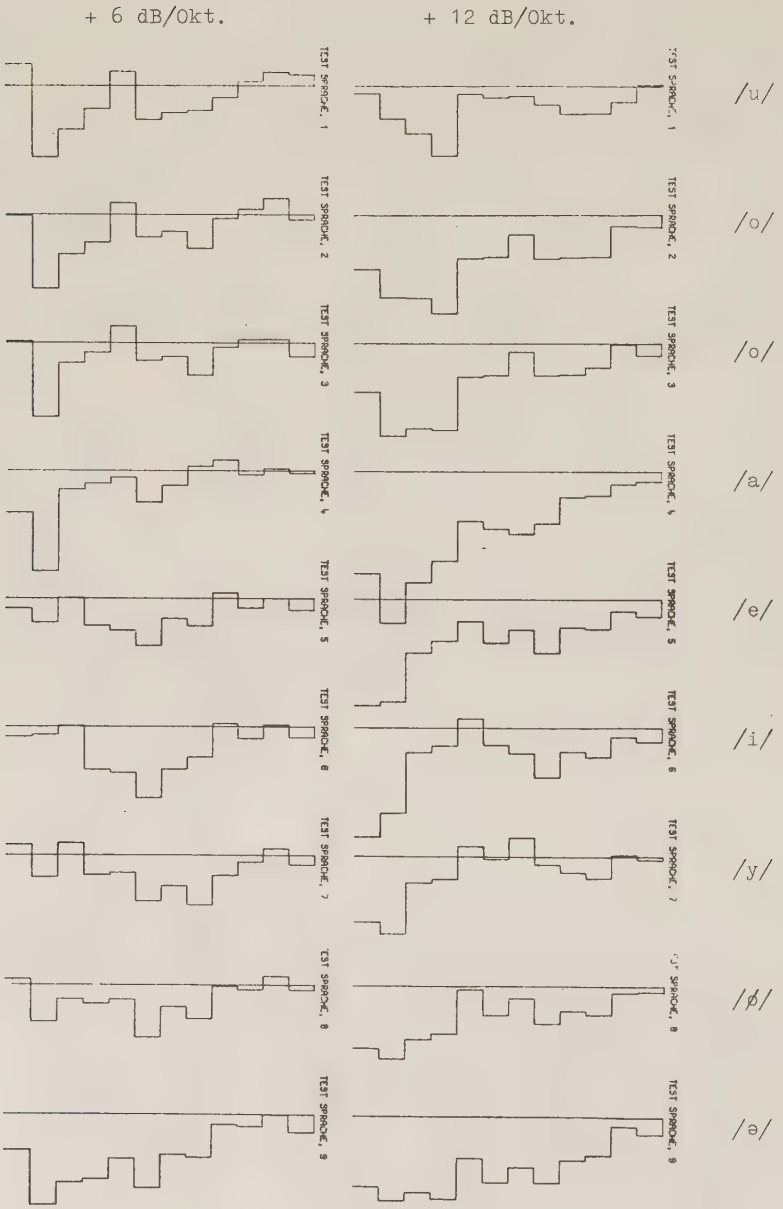


Fig. 17 b: Wie a, natürliche Vokale. Links einmal, rechts zweimal differenzierter Schalldruck

Forts.

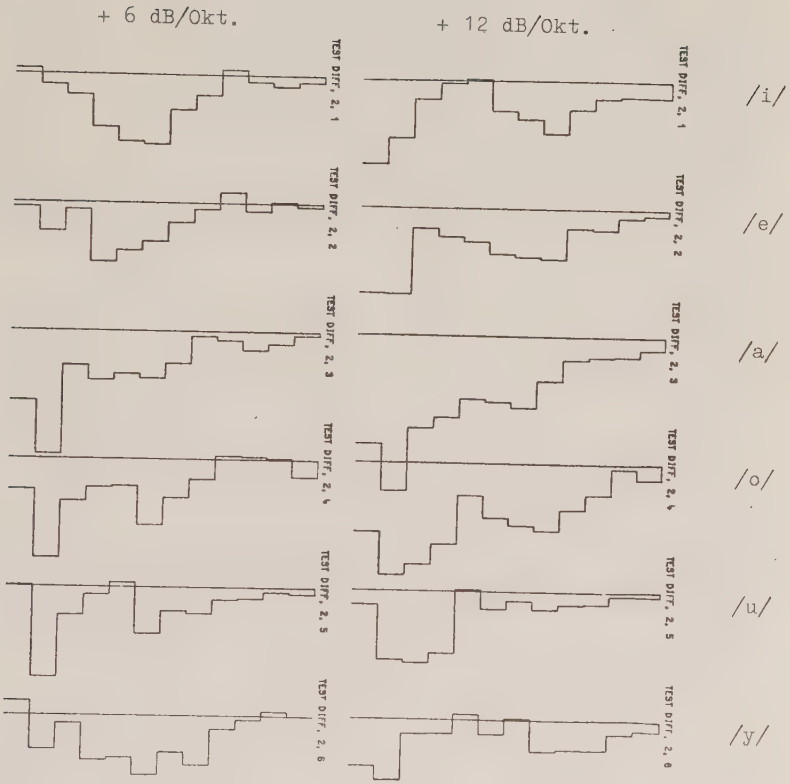


Fig. 17 b, Fortsetzung

stigste Höhenanhebung scheint zwischen 6 und 12 dB/Okt. zu liegen, was z.B. die Ergebnisse für /i/ besonders gut zeigen. Der Frequenzgang des Glottisspektrums ist jedoch auch für einen Sprecher nicht konstant; er hängt z.B. von der Lautstärke ab und schwankt etwa zwischen -11 und -17 dB/Okt. [16].

Die aus der ADPCM-Technik bekannte Neigung der Reflexionsfaktoren r_n (= PARCOR-Koeffizienten), nur für kleine n groß zu sein und im Vorzeichen zu alternieren [62], tritt hier (insbesondere bei 12 dB/Okt. Anhebung) nicht mehr immer deutlich auf. Sie beruht nämlich offenbar auf schwachen oder fehlenden hohen Frequenzen. Nehmen wir etwa ein Signal an, für das der nach dem Autokorrelationsverfahren (A.3) ermittelte Prädiktor das exakte (reziproke) Spektrum darstellt, und integrieren es einmal näherungsweise durch Filterung mit $1/(1 - Kz^{-1})$, $K \lessapprox 1$, was einer fehlenden Höhenanhebung bei wirklicher Sprache nahekommt. Dadurch erhält der Prädiktor einen zusätzlichen Faktor $1 - Kz^{-1}$, die neuen Prädiktorkoeffizienten sind $b_k^n = a_k^{n-1} - Ka_{k-1}^{n-1} \approx a_k^{n-1} - a_{k-1}^{n-1}$; die b_k^n erfüllen für $K \to 1$ ebenfalls (A:11). Der erste Reflexionsfaktor wird also groß und negativ ($b_1^1 = -K$), die übrige Rohrform wird verschoben und invertiert ($b_n^n \approx -a_{n-1}^{n-1}$, $S_n \mapsto S_{n-1}^{-1}$ in Atals Auffassung), ein erneuter Beweis für den enormen Frequenzgangeinfluß. Wenn die r_n im Vorzeichen alternieren, löschen sich die reflektierten Wellen mit den hinlaufenden teilweise aus, wenn man sich der halben Abtastfrequenz nähert (vgl. "verbotene Bänder" bei Kristallwellen); fehlende Energie am oberen Bandende kann also auf diese Weise dargestellt werden.

Um unabhängig vom unbekannten Glottisspektrum, vom Gesamtfrequenzgang und von der oben erwähnten Formantverschmierung (2.1, 5.1.2) zu werden, wurde versucht, statt von der AKF eines längeren Signalabschnitts (nach A.3) von einem Prädiktor A_p auszugehen, der die Prädiktionsfehler-Quadratsumme (A:4) nur auf einem Glottisverschlußintervall (nach A.2) minimiert ($w_k = 1$ dort, $= 0$ sonst). Dort sind die Formanten konstant und die Annahme einer harten Glottis erfüllt, und wenn H_u näherungsweise nur Pole hat, so sollte $H_u \lessapprox 1/A_p$ sein. Es wurden 20 Abtastwerte = 2,3 ms benutzt. Die Kürze des Signalsegments macht das Verfahren leider anfälliger gegen Ungenauigkeiten; gelegentlich können (im Gegensatz zum Verfahren A.3) negative

Bandbreiten herauskommen. Nach (A:13) gebildete Prädiktorpolynome niedrigeren Grades sind auch nicht mehr identisch mit den aus der Autokovarianzmatrix nach (A:6) für niedrigere p berechneten Prädiktoren, so daß die erhaltene Querschnittsfunktion vom Prädiktorgrad abhängt. Ist p größer als die doppelte Zahl der Formanten (= Rohrlänge/Segmentlänge), so treten weitere Pole auf, die das Spektrum des nach Inversfilterung unter Verwendung der Formanten verbleibenden Fehlersignals wiedergeben. Solche Zusatzpole können jedoch, außer wenn sie stark gedämpft sind, die Querschnittsfunktion unvorhersehbar erheblich verfälschen. Selbst wenn aber $1/A_p$ die exakte Übertragungsfunktion wäre, käme aufgrund der Abweichungen zwischen Atals und Wakitas Modellen und der Wirklichkeit nicht die wahre Querschnittsfunktion heraus.

Einige Beispiele wurden durchgerechnet. Für das homogene, ohmsch abgeschlossene Rohr ohne Querzweig ist das Kovarianzverfahren, auf das Intervall vom 5. bis 25. Wert der Impulsantwort angewandt, wirklich unempfindlich gegen Differentiation. Einführung des Querzweiges liefert für das richtige p = 9 etwa dasselbe wie das Korrelationsverfahren (vgl. Fig. 15, 3), für p = 10 wird das Ergebnis besser, abgesehen von der Sprunghöhe am Mund; für p > 10 wird die Kovarianzmatrix singulär (Fig. 15, 4). Überlagert man dem Signal Rauschen (-35 dB), so ähnelt das Ergebnis auch für p > 9 wieder mehr dem des Korrelationsverfahrens, der Gesamtfrequenzgang des Prädiktorfilters ist jedoch verschieden (Fig. 15, 5).

Fig. 18 zeigt wieder Ergebnisse für (a) synthetische Vokale nach 4.1 (p = 8, 10, 12) und (b) für natürliche Vokale (p = 10). Der Prädiktionsfehler wurde auf einem Teil des Glottisverschlußintervalls minimiert (zur Bestimmung des Intervallanfangs bei natürlicher Sprache siehe 4.3); auch bei den synthetischen Vokalen war jedoch das Eingangssignal wegen der Tiefpaßfilterung dort nicht exakt Null. Für die synthetischen Vokale sind die Ergebnisse besser als nach dem Korrelationsverfahren, auch der Mundverschluß bei /u/ kommt etwas besser heraus. Atals Deutung ist bei offenen Vokalen und nicht zu kleinem p klar im Vorteil, sonst jedoch ist die Entscheidung unklar. Die Ergebnisse für natürliche Vokale ähneln denen in Fig. 17 b oft wenig; mal kommen sie denen für einmalige, mal denen für zweimalige Differentiation näher, was alles auf Ab-

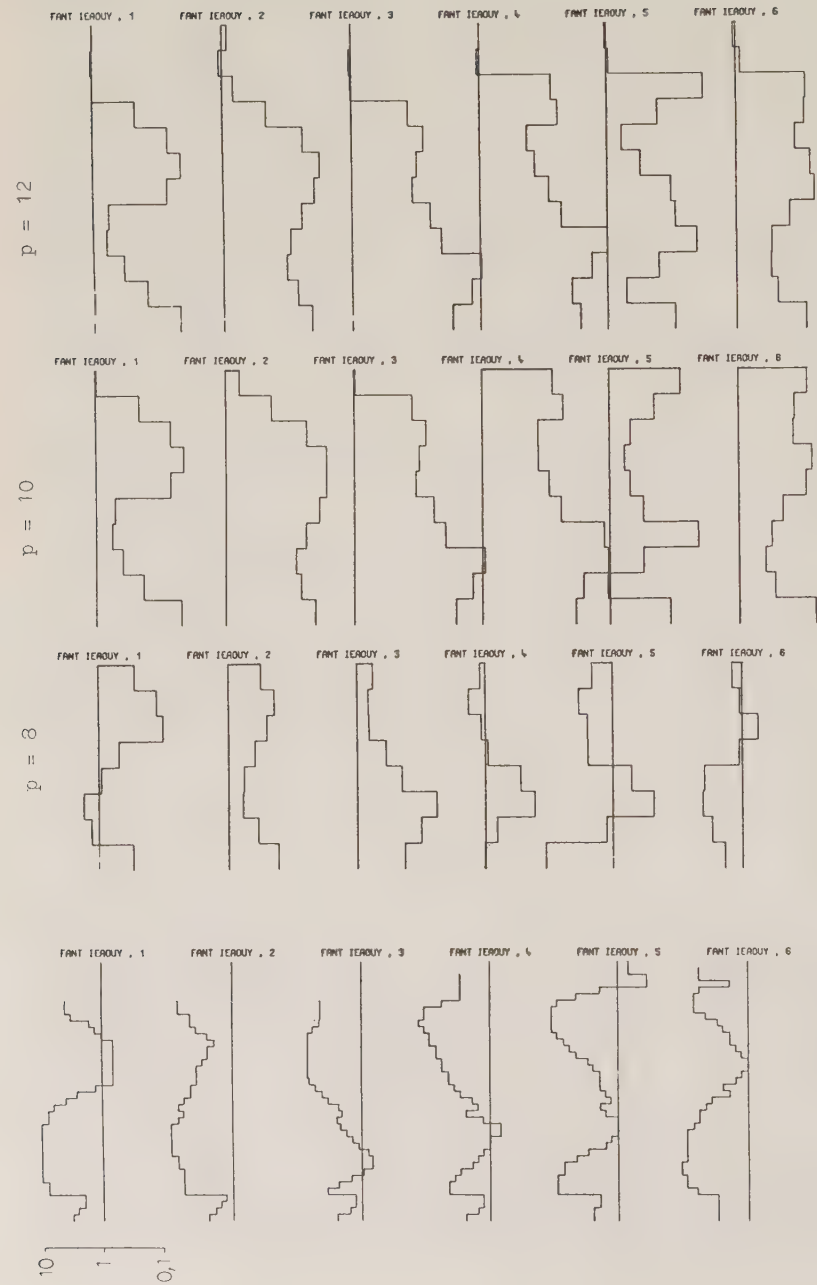


Fig. 18 a: Bestimmung der Querschnittsfunktion aus Prädiktor im Glottisverschlußintervall, Grad p = 8, 10, 12; d = 2 cm. Fant-Vokale (links Originale, d = 0,5 cm). Darstellung nach Atal; nach Wakita Kurven um 180° drehen!

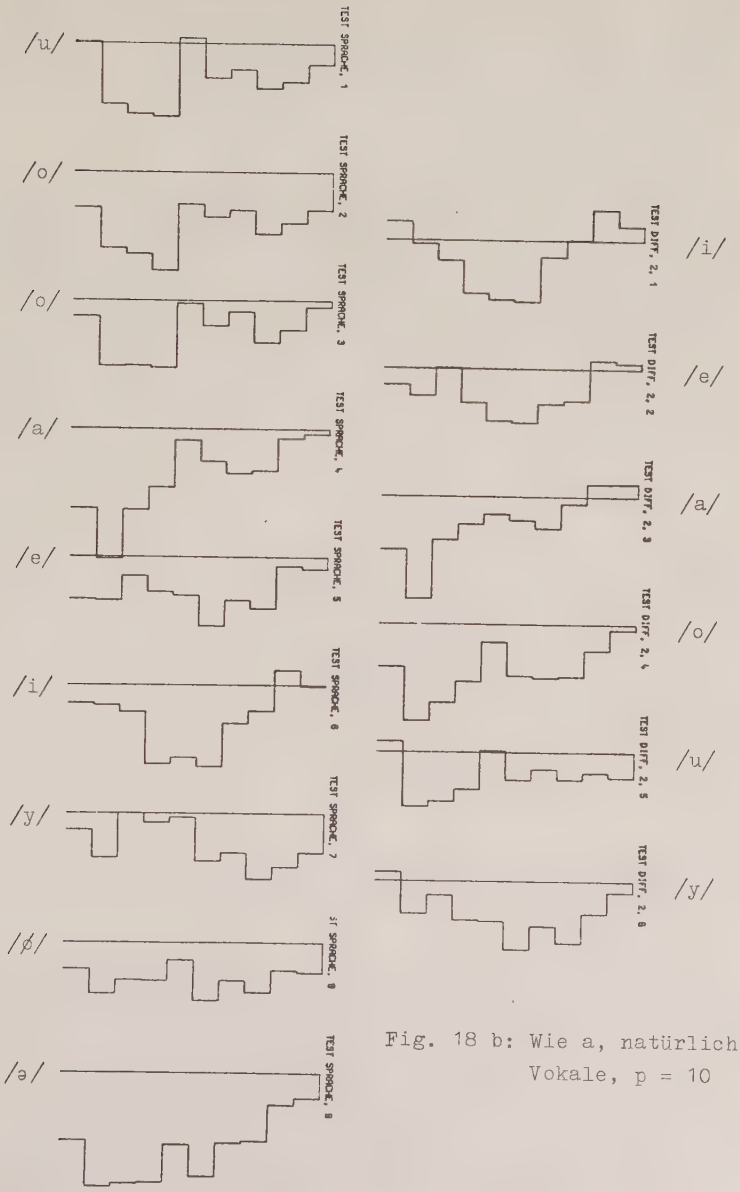


Fig. 18 b: Wie a, natürliche Vokale, p = 10

weichungen vom Nur-Pole-Modell mit reell abgeschlossenem, verlustlosem Rohr, Schwankungen des Glottisspektrums oder nicht hinreichend gute Ausprägung des Glottisverschlußintervalls zurückgehen dürfte. Enge Mundöffnungen kommen wieder nur in Wakitas Deutung heraus (Kurven um 180° drehen!).

5.2.4. Folgerung

Zusammenfassend muß die Anwendbarkeit von direkten Rechenverfahren zur Bestimmung der Querschnittsfunktion aus dem Sprachsignal ohne a-priori-Einschränkung möglicher Lösungen pessimistisch beurteilt werden, da - neben den in 5.1. dargestellten Schwierigkeiten - die Lösung viel zu sehr von der genauen Meßbarkeit der relevanten akustischen Größen, insbesondere der Formantbandbreiten, unter Elimination des Glottisspektrums abhängt. Auch die zugrunde liegenden physikalischen Modellvorstellungen, wie Annahmen über Dämpfungsursachen und Wandimpedanz, haben einen äußerst starken Einfluß auf die Lösung.

Nach allerneuesten Ergebnissen scheint dennoch Wakitas Verfahren für gewisse praktische Zwecke, wie visuelle Hilfen für Taubstumme [1], geeignet zu sein; auch wurde es mit komplizierteren Methoden zum Ausgleich des Frequenzganges kombiniert [63]. Die Verfahren sind aber wohl mehr durch das im Groben vernünftige Aussehen der Ergebnisse als durch die Theorie gerechtfertigt; auch eine persönliche Diskussion mit den Autoren brachte keine weitere Klärung.

5.3. Versuche mit iterativer Modellanpassung

5.3.1. Grundlegendes

Der Versuch, die Querschnittsfunktion mittels direkter Algorithmen zu bestimmen, setzt die Anwendbarkeit einfacher physikalischer Modellvorstellungen voraus. Wie wir gezeigt haben, können schon kleine Abweichungen zwischen Modell und Wirklichkeit zu völlig falschen Ergebnissen führen. Kompliziertere Modelle, die etwa Glottisspektrum, Wandimpedanz und Viskositätsdämpfung berücksichtigen, werden aber i.a. nicht auf direkte Algorithmen führen, ganz abgesehen davon, daß die-

se zusätzlichen Größen i.a. ebenso unbekannt sind wie die Querschnittsfunktion. Lediglich für zwei bestimmte Verteilungen von zusätzlichen Längs- und Querimpedanzen geben Sondhi und Gopinath [46, 34] Rechenverfahren an, die aber schwierig auszuführen sind. Größere Freiheiten in der Verwendung komplizierterer Stimmkanalmodelle bietet die iterative Bestimmung der Unbekannten unter Minimierung irgendeines Fehlermaßes zwischen gemessenen und vom Modell vorhergesagten akustischen Größen. Als Unbekannte kann man (bei Verwendung der Stufenleitung) im einfachsten Fall die Querschnitte S_n oder einfache Funktionen davon wählen; dann muß die Segmentierung aber wie bei den direkten Verfahren sehr grob sein, wenn nicht künstlich hochfrequente Schallanteile hinzugefügt werden (vgl. 5.1). Dies läßt sich jedoch jetzt umgehen, da die gesamte Querschnittsfunktion durch ein Modell mit wenigen Parametern beschrieben werden kann (Fig. 19). Damit ergibt sich die Mög-

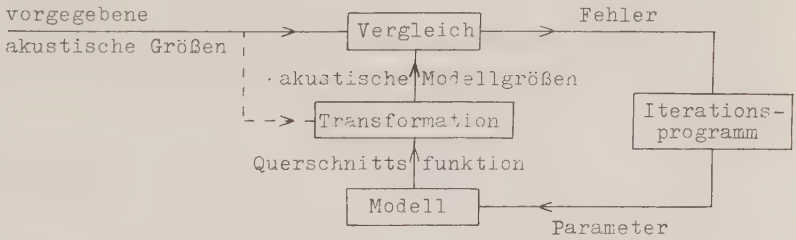


Fig. 19: Iterative Modellanpassung

lichkeit einer beliebig feinen Rohrsegmentierung, d.h. einer besseren Approximation der wirklichen Verhältnisse, sowie die leichte a-priori-Einschränkbarkeit der Menge möglicher Querschnittsfunktionen, indem man etwa ein Artikulatormodell zugrunde legt. Ein Beispiel ist die Kosinusreihen-Darstellung (5:1), wo nur die ersten Fourierkoeffizienten bestimmt werden, evtl. sogar nur die mit ungerader Nummer. Da dies Modell eine relativ einfache Beziehung zwischen Modellparametern und akustischen Größen aufweist (vgl. (5:2)), soll es auch in dieser Arbeit verwendet werden. Es hat den Nachteil, daß die Menge der möglichen Lösungen, außer durch willkürliche Annahmen wie (5:6) und abgesehen von räumlicher Tiefpaßfilterung durch Benutzung nur endlich vieler Fourierkoeffizienten, nicht eingeschränkt ist. Daher soll auch ein kompliziertes Artikulator-

modell nach Mermelstein und Maeda mit physiologisch sinnvollen Parametern verwendet werden. Die Übertragungsfunktion wird grundsätzlich gemäß (3:5) für eine Stufenleitung aus 0,5-cm-Segmenten konstruiert; Strahlungsdämpfung und Wandimpedanz werden berücksichtigt, jedoch keine weiteren Dämpfungsquellen. Die Wandimpedanz kennt man zwar nur sehr grob; da ihr genauer Wert aber nicht so wichtig zu sein scheint, wurden wieder die Werte nach Fig. 6 (a) benutzt.

Ein anderes Problem besteht darin, welche akustischen Größen für die Bestimmung der Querschnittsfunktion verwendet werden sollen. In Frage kommen etwa Abtastwerte des Spektrums, die Formantfrequenzen und -bandbreiten, ein Stück der Autokorrelationsfunktion. Da bei Benutzung der Formantfrequenzen und -bandbreiten der Übertragungsfunktion des Stimmkanalmodells deren Bestimmung bei jedem Iterationsschritt neu erfolgen müßte (sogar mehrmals zur Bildung partieller Ableitungen), was rechenzeitmäßig extrem aufwendig und mit zu großen Rechenungenauigkeiten behaftet wäre, wurde dieser Weg nicht weiter verfolgt. Einfach dagegen ist eine Anpassung des Spektrums, wobei für wirkliche Sprache jedoch die Richtigkeit des Gesamtfrequenzganges und die Elimination des Glottiswellen-Einflusses Schwierigkeiten macht.

Da die Spektrumsbestimmung in der Praxis vorteilhaft mittels linearer Prädiktion geschieht und da die Prädiktion einer Inversfilterung entspricht, liegt eine andere Möglichkeit der Wahl der anzupassenden Größen nahe: Man filtert die Sprache mit einer aus der Übertragungsfunktion des Modells konstruierten inversen Übertragungsfunktion auf einem Glottisverschlußintervall und minimiert das resultierende Signal, das ja dem (verschwindenden) Glottissignal entspricht. Hierbei ist man wieder unabhängig von der unbekannten Gestalt der Glottiswelle. Die Versuche mit den beiden Verfahren werden weiter unten genauer beschrieben.

5.3.2. Das Iterationsprogramm

Die Querschnittsfunktion des Stimmkanals werde durch die Parameter $\underline{b} = (b_1,\ldots,b_k)^T \in \mathbb{R}^k$ beschrieben. Die gemessenen akustischen Größen seien $\underline{y} = (y_1,\ldots,y_n)^T \in \mathbb{R}^n$. Das Stimmkanalmodell liefert für den Parametersatz $\underline{b}$ die $\underline{y}$ entsprechenden

Werte $\hat{\underline{y}} = \underline{F}(\underline{b})$, $\underline{F}: \mathbb{R}^k \mapsto \mathbb{R}^n$. Dann ist ein Wert von $\underline{b}$ gesucht, so daß

$$\Phi := |\underline{y} - \underline{F}(\underline{b})|^2 = \sum_{i=1}^{n} (y_i - F_i(\underline{b}))^2 \qquad (5:9)$$

minimal wird. In jedem Fall muß die Zahl n der anzupassenden Werte mindestens gleich der Zahl k der Unbekannten sein, wenn die Lösung lokal eindeutig sein soll.

Zur Lösung dieses Problems wurde ein vorhandenes, sehr vielseitig verwendbares Anpaßprogramm benutzt, das von einem Iterationsverfahren nach Marquardt [64] Gebrauch macht. Sei bei irgendeinem Iterationsschritt

$$\underline{F}_i = \partial \underline{F}/\partial b_i \in \mathbb{R}^n, \quad \underline{P} = \left(\frac{\underline{F}_1}{|\underline{F}_1|}, \ldots, \frac{\underline{F}_k}{|\underline{F}_k|}\right) \in \mathbb{R}^{nxk},$$
$$\underline{g} = (\underline{y} - \underline{F}(\underline{b}))^T \underline{P} \in \mathbb{R}^k; \quad \underline{A} = \underline{P}^T \underline{P} \in \mathbb{R}^{kxk} \qquad (5:10)$$

ist die Korrelationsmatrix der partiellen Ableitungen. Dann erhält man $\Delta\underline{b}$, die Korrektur zu $\underline{b}$, aus dem linearen Gleichungssystem

$$\lambda \underline{x} + \underline{A}\underline{x} = \underline{g}, \qquad \Delta b_i = x_i/|\underline{F}_i|, \; i = 1,\ldots,k. \qquad (5:11)$$

Dies ist ein Kompromiß zwischen Gradienten- und Newton-Verfahren, entsprechend dem 1. bzw. 2. Term auf der linken Seite; λ wird nach gewissen Kriterien geeignet ($\approx 10^{-8}$) gewählt, groß z.B. bei den ersten Iterationsschritten oder nicht mehr abnehmendem Fehlermaß Φ (5:9). So wird die Sicherheit des Gradientenverfahrens bei ungünstigen Anfangswerten mit der schnelleren Konvergenz des Newton-Verfahrens nahe der Lösung vereint. λ verhindert auch Schwierigkeiten, wenn $\underline{A}$ durch Rechenungenauigkeiten singulär oder indefinit ist. Das Verfahren wird in der Regel als konvergiert betrachtet, wenn

$$\frac{|\Delta b_j|}{|b_j| + \tau} < \varepsilon, \; j = 1,\ldots,k \quad (\text{"}\varepsilon\text{-Test"}), \qquad (5:12)$$

τ und ε lassen sich vorgeben. Kleines τ ermöglicht Abfrage auf die relativen Parameterkorrekturen, großes τ ($\gg |b_j|_{max}$) auf die absoluten. Ein grobes Flußdiagramm des Programms zeigt Fig. 20. Da in unserem Fall die partiellen Ableitungen $\underline{F}_i$ praktisch nicht analytisch angebbar sind, werden sie als Differenzenquotienten berechnet, wobei

$$\delta b_i = D \cdot b_i \text{ (wenn Null, wird } \delta b_i = D \text{ gesetzt).} \qquad (5:13)$$

Zu kleine δb_i führen wegen Rechenungenauigkeiten zu unsinnigen Ergebnissen; beim Rechnen auf einer Honeywell H632, bei der die Mantisse der Gleitkommazahlen nur aus 6 Hexadezimalstellen besteht, wurde zunächst D = 1/256, später sogar = 1/64 gesetzt.

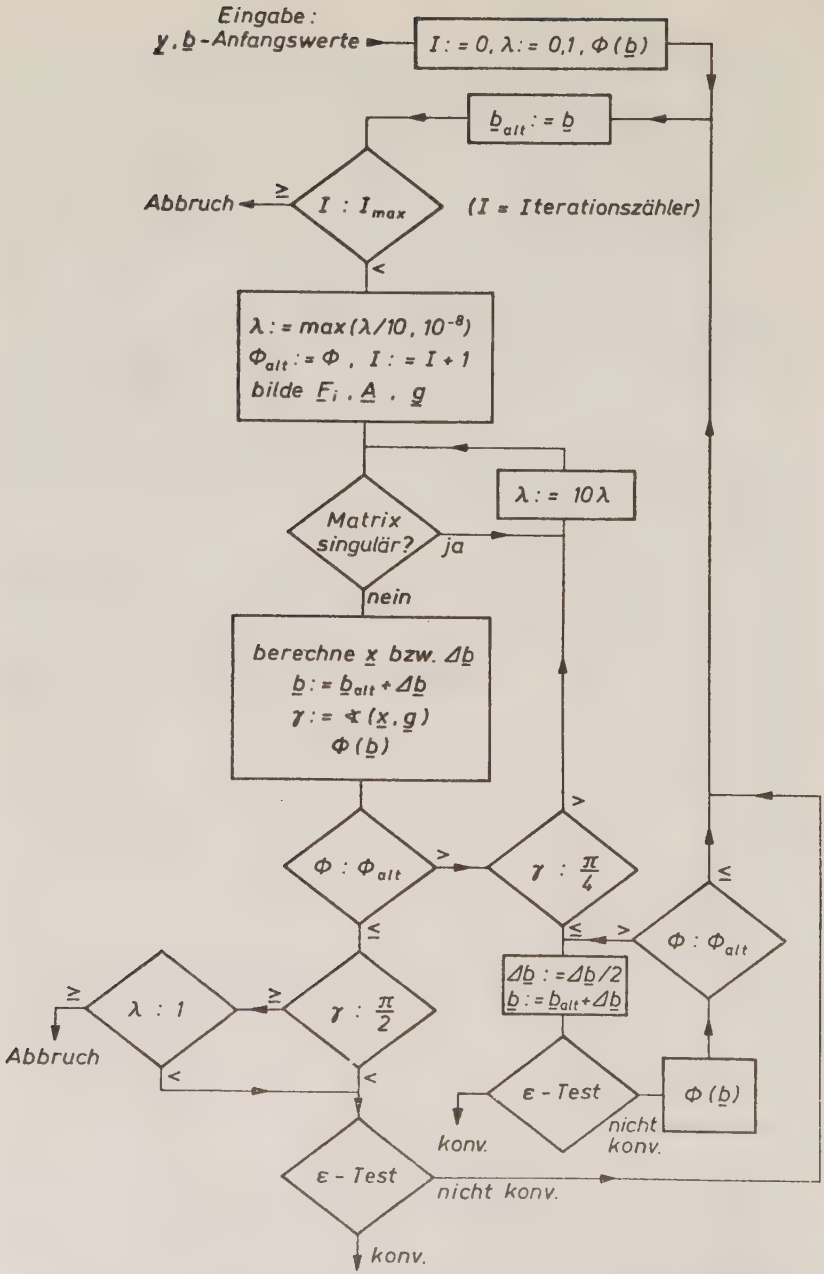


Fig. 20: Flußdiagramm des Marquardt-Iter[illegible]onsprogramms

Bei sehr unterschiedlicher Bedeutung der einzelnen b_j, etwa in einem Artikulatormodell, sollte ihr Wertebereich durch geeignete Skalierung so angeglichen werden, daß die Verwendung von für alle j gleichen ε-, τ- und D-Werten Sinn hat. Das Programm erlaubt es übrigens, einzelne Parameter bei der Iteration festzuhalten.

5.3.3. Die benutzten Modelle für die Querschnittsfunktion

5.3.3.1. Das Kosinusreihen-Modell

Die erste Version des Modells ist ein diskretes Analogon zu dem von Mermelstein und Schroeder verwendeten, (5:1):

$$S_n = \exp \sum_{j=0}^{k-1} b_j \cos \frac{\pi j(n-1)}{N-1} \quad , \; n = 1,\ldots,N. \qquad (5:14)$$

Die Zahl N der Segmente und damit die Länge des Stimmkanals muß vorgegeben werden. Da dieses Modell den immer vorhandenen Querschnittssprung am Ende des Larynxrohres nicht wiedergibt, wurde später ein solches Rohr angesetzt. Bei einer Segmentlänge von d = 0,5 cm bestand es aus vier Segmenten; nach Fants Querschnittsfunktionen wurde der Querschnitt an der Glottis als 3 cm^2 angesetzt und die Querschnittszunahme am Ende des Larynxrohres als 1 : 4. Mit linearer Interpolation dazwischen wird nun

$$S_n = \begin{cases} \exp \sum_{j=0}^{k-1} b_j \cos \frac{\pi j(n-5)}{N-5} \quad , \; n = 5,\ldots,N, \\ 3 + (S_5/4 - 3)\frac{n - 1/2}{4} \quad , \; n = 1,\ldots,4. \end{cases} \qquad (5:15)$$

Das Modell hat den Vorteil, wenig Rechenzeit zu benötigen; solange die Rohrlänge N gleich bleibt, können die Kosinuswerte in einer Tabelle abgelegt werden. Außerdem erscheint die exponentielle Abhängigkeit der S_n von den Parametern b vorteilhaft, da die S_n dann automatisch > 0 sind und da die Übertragungsfunktion im wesentlichen nur von Querschnittsverhältnissen bestimmt wird.

Achtung: Die b_j sind hier von 0 bis k-1 statt von 1 bis k numeriert!

5.3.3.2. Das Artikulatormodell

Die Verwendung eines "physiologischen" Artikulatormodells hat den Vorteil, daß aus der Menge möglicher Querschnittsfunktionen von vornherein diejenigen ausgeschlossen werden, die

von Bau und Funktion des Sprechapparates her nicht vorkommen können. Daher braucht man evtl. weniger akustische Information, um die richtige Lösung zu finden. Nachteile sind, daß der Zusammenhang zwischen Modellparametern und Querschnittsfunktion bzw. Übertragungsfunktion äußerst kompliziert wird und daß man über die wirklichen Verhältnisse im Stimmkanal nicht mehr erfährt, als bereits im Modell vorgesehen ist.

Ein äußerst einfaches, aber noch nicht sehr wirklichkeitsnahes Modell stammt von Stevens und House [65]und wurde unlängst von Atal verbessert [66]. Die anatomischen Verhältnisse besser abbildende Modelle haben Coker [9], Lindblom und Sundberg [67] und Mermelstein und Maeda [10] konstruiert. Wichtig ist, daß bereits vernünftige Annahmen über den Zusammenhang zwischen dem vom Modell primär gegebenen sagittalen Abstand Zunge-Rückwand und dem Querschnitt des Stimmkanals gemacht werden. Die Ergebnisse der akustischen Bestimmung der Querschnittsfunktion mittels Artikulatormodell können also zur Klärung dieses in 1. angesprochenen Problems nicht viel beitragen. Alle diese Modelle machen komplizierte Koordinatentransformationen erforderlich, die in den Veröffentlichungen nicht explizit angegeben sind. Hier soll eine Abart einer älteren, unveröffentlichten Version des Mermelstein-Maeda-Modells benutzt werden, die vor Erscheinen von [10] aus Abbildungen und unvollständigen Mitteilungen (persönliche Korrespondenz) von mir konstruiert wurde. Die Autoren ermittelten ihr Modell durch gleichzeitige Anpassung des Sagittalschnitts und der Formantfrequenzen unter Verwendung von synchronen Röntgen- und Sprachaufnahmen. Die Modellparameter geben die Unterkieferstellung, Lage des Zungenkörpers, Zungenspitzenstellung, Lippenrundung, Stellung des Velums und Lage des Zungenbeins an. Dieser Parametersatz ist nichtminimal, also akustisch nicht eindeutig rekonstruierbar (z.B. ist die Lage des Zungenkörpers relativ zum Unterkiefer angegeben, so daß sich Änderungen der Kieferstellung durch Änderungen der Zungenparameter teilweise kompensieren lassen). Die Unterkieferstellung bestimmt mit der Lippenrundung die genaue Lippenform.

Für die Zwecke dieser Arbeit, in der es nur um Vokale geht, wurden nur fünf Parameter betrachtet: Unterkieferstellung, Lage des Zungenkörpers (2 Koordinaten), Lippenrundung, Gesamtmaßstab. Der letztere wurde fast immer als gegeben be-

trachtet, wie die Länge im Kosinusreihen-Modell. Die Unterkieferstellung läßt sich nötigenfalls durch direkte Messungen ermitteln oder schätzen, so daß noch eine Unbekannte wegfällt (ein Vorteil, der mit zur Auswahl dieses Modells beigetragen hatte). Die verbleibenden drei bis vier Parameter sollten bei Kenntnis von drei bis vier Formanten bestimmbar sein, evtl. selbst bei ungenau bekannten Bandbreiten.

Das Modell besteht aus einem festen und einem mit dem Unterkiefer drehbaren Teil sowie den Lippen. Im festen Koordinatensystem ist der Verlauf der Stimmkanal-Rückwand definiert (Fig. 21 oben). Die Längenkoordinate x wird im Rachen ($x < 0$) und im Mundvorderteil ($x > 6$) cartesisch, dazwischen als Winkel (Einheit $\pi/10 = 18^{\circ}$) gemessen, siehe gestrichelte Linie. Für $1{,}6 < x \leq 5$ nimmt die Rückwandkoordinate linear von y_3 auf y_4 ab, bleibt dann bis $x = 6$ konstant; für $6 < x \leq 8$ ist $y(x)$ ein Polynom 3. Grades mit Extrema bei $(6,y_4)$ und $(8,y_5)$, danach bis $x = 9$ eine Gerade mit Steigung $-1{,}5$. Der Punkt J ist der Drehpunkt des Unterkiefers, dessen Stellung durch einen Winkel α_J (im Bogenmaß) gemessen wird. Am Unterkiefer ist ein Koordinatensystem befestigt, das dem obigen gleicht und für $\alpha_J = 0$ mit diesem zusammenfällt. Im bewegten Koordinatensystem wird die Zungenform $t(x)$ beschrieben (Fig. 21 Mitte); für $x \leq 6$ ist $t(x) = s(x) + d(x)$,

$$s(x) = \begin{cases} 2{,}2\left(\sqrt{(x+6{,}75)/3} - 1/2\right), & x < 0, \\ 2{,}2 & , \; x \geq 0, \end{cases}$$

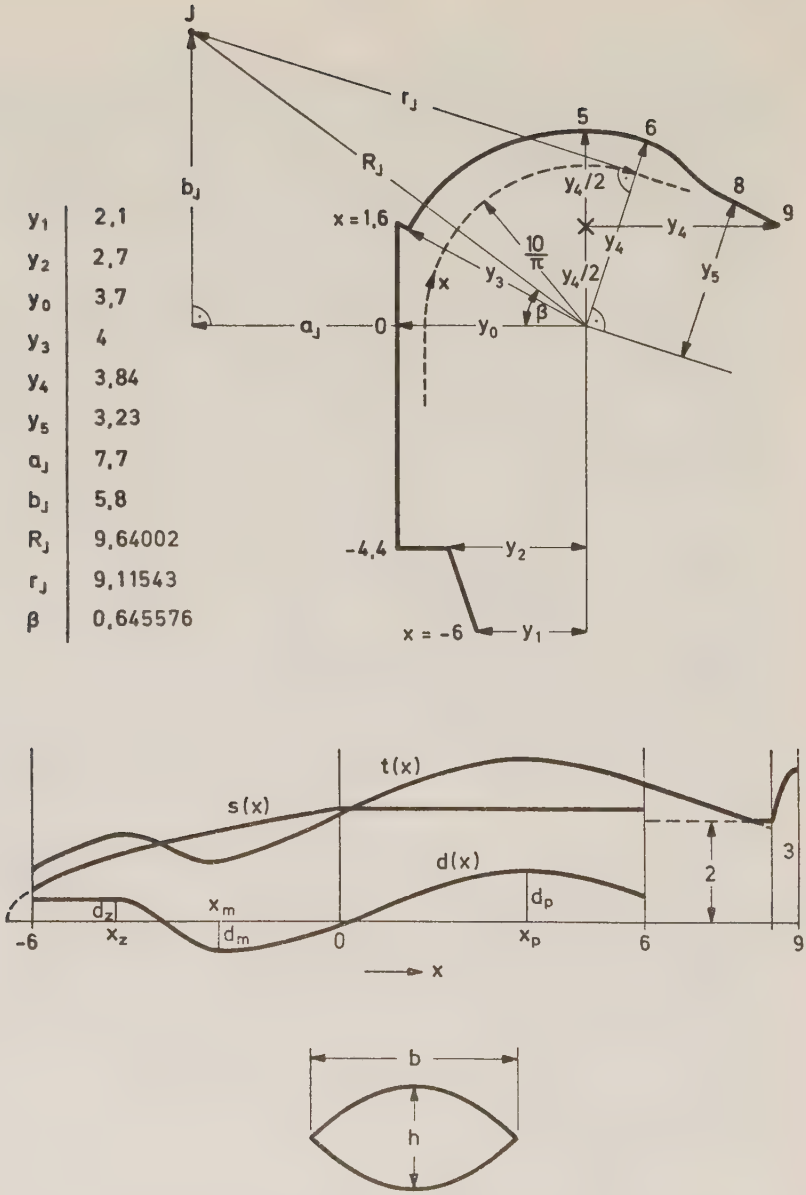
$d(x)$ ist durch drei Punkte (x_z,d_z), (x_m,d_m), (x_p,d_p) festgelegt: $d(x) = d_z$ für $x \leq x_z$; zwischen x_z und x_m und zwischen x_m und 6 wird jeweils mit Kosinus-Bögen mit Extrema bei x_z, x_m, x_p interpoliert. x_z wird $= -4{,}4$ gesetzt und d_z in Abhängigkeit von α_J so gewählt, daß der Abstand zwischen Zunge und Rückwand an der Glottis 1 beträgt. Zwischen den anderen Punkten bestehen die Beziehungen

$$x_m = 1{,}1x_p + 1{,}3d_p - 0{,}55x_pd_p - 4{,}7,$$
$$d_m = 1{,}2 - 0{,}55d_p - 0{,}18x_p$$

und die Einschränkungen

$$0 \leq x_p \leq 6, \quad d_m \leq d_p \leq \begin{cases} 1{,}3034 + 0{,}86x_p & , \; x_p < 0{,}81 \\ 2 & , \; x_p \geq 0{,}81 \end{cases} \quad (5{:}16)$$

(hierdurch wird u.a. ausgeschlossen, daß $d(x)$ ein zweites Maximum zwischen 0 und 6 annimmt). - Für $6 \leq x \leq 8{,}5$ wird $t(x)$ als Tangente fortgesetzt, aber auf ≥ 2 begrenzt. Danach erfolgt ein parabolischer Übergang auf die unteren Schneidezäh-



y_1	2,1
y_2	2,7
y_0	3,7
y_3	4
y_4	3,84
y_5	3,23
a_J	7,7
b_J	5,8
R_J	9,64002
r_J	9,11543
β	0,645576

Fig. 21: Artikulatormodell nach Mermelstein und Maeda; oben: festes Koordinatensystem und Rückwand; Mitte: bewegtes Koordinatensystem (gestreckt) mit Zungenform; unten: Mundform

ne, $t(9) = 3$. - Die Koordinatentransformationen zwischen festem und bewegtem System werden nur grob angenähert (exakt erhielte man z.T. transzendente Gleichungen). In beiden Systemen wird die gleiche x-Koordinate verwendet und nur die Lage von x_m und x_p in Abhängigkeit von α_J näherungsweise transformiert:

$$x_m \mapsto x_m + \alpha_J(t(x_m) - a_J),$$

$$x_p \mapsto x_p - \alpha_J \frac{10}{\pi}\left(\frac{R_J}{t(x_p)} \cos\left(\frac{\pi}{10}x_p - \beta\right) - 1\right).$$

Es werden nun an Stellen x_n die Ordinatenwerte w_n der Rückwand und t_n der Zunge im festen System bestimmt; dabei sind die x_n so zu finden, daß auf der Mittellinie $m_n = (w_n + t_n)/2$ des Stimmkanals gleiche Abschnitte $\Delta x = d/s$ (d gewünschte Segmentlänge der Stufenleitung, s Skalenfaktor) abgeteilt werden. Für $x_n + \Delta x \leq 1{,}6$ wird einfach $x_{n+1} = x_n + \Delta x$ gesetzt. Für $1{,}6 < x_n + \Delta x \leq 9$ wird dagegen ein geneigter Verlauf der Mittellinie berücksichtigt: $x_{n+1} = x_n + u\Delta x$, $t_{n+1} = t_n + u\Delta t$, $w_{n+1} = w_n + u\Delta w$

mit $u = \begin{cases} \left(\left((m_n+\Delta m)\frac{\pi}{10}\right)^2 + (\Delta m/\Delta x)^2\right)^{-1/2}, & x < 6-\Delta x, \\ \left(1 + (\Delta m/\Delta x)^2\right)^{-1/2}, & x \geq 6-\Delta x, \end{cases}$

wo $t_n + \Delta t$ der zu $x_n + \Delta x$ berechnete t-Wert ist (analog Δw, Δm). Die t-Werte ergeben sich aus den im bewegten System definierten $t(x)$ durch folgende genäherte Transformation:

$$t = t(x) - \begin{cases} \alpha_J(x - b_J), & x \leq 0, \\ \alpha_J R_J \sin\left(\frac{\pi}{10}x - \beta\right), & 0 < x < 6, \\ \alpha_J(r_J + x - 6), & 6 \leq x. \end{cases}$$

Bei Erreichen von $x_n + \Delta x > 9$ wird zunächst aufgehört, $M := n$ gesetzt und (nur für Plot-Zwecke) $t_{M+1} = 3 - \alpha_J(r_J + 3)$ definiert. Die sagittalen Abstände $d_n = (w_n - t_n)s$ werden nun in Querschnittswerte umgerechnet:

$$S_n = \begin{cases} v_n d_n, & x_n \leq 1{,}6, \\ 2{,}094\, d_n \sqrt{|d_n|}, & x_n > 1{,}6, \end{cases} \qquad v_n := \begin{cases} 1{,}178, & x_n \leq -4{,}4, \\ 1{,}571 + 0{,}785\frac{x_n + 4}{3}, & -4{,}4 < x_n \leq -1, \\ 2{,}356, & -1 < x_n \leq 1{,}6 \end{cases}$$

(im Pharynx ist der Querschnitt elliptisch mit festen Querabmessungen angenommen, was nicht wirklich stimmt), ferner wird S_n auf $\geq 0{,}01\ cm^2$ eingeschränkt, da die d_n negativ werden können.

Nun muß noch der Lippenabschnitt angesetzt werden, dessen Abmessungen durch Kieferstellung α_J und Lippenrundung r gegeben sind. Als Lippenform wurden Parabelbögen angesetzt (Fig.

21 unten). Für die Höhe, Breite und Länge des Lippenrohres gilt (in cm):

$$h = 7{,}72\alpha_J + 0{,}52 - r,$$
$$b = -37{,}5\alpha_J + 5{,}2h + 1{,}2,$$
$$l = 7{,}4\alpha_J - 0{,}7h + 1{,}72,$$

für die Fläche $S = \frac{2}{3}bh = (-25\alpha_J + 0{,}8 + 3{,}46667h)h$. Kommt $h \leq 0$ oder $S < 0{,}01\ cm^2$ heraus, wird $S = 0{,}01\ cm^2$ gesetzt und auch h und r entsprechend verändert. Das Lippenrohr wird aus weiteren Δx-Segmenten bei $x_{M+1},\ldots,x_N$ aufgebaut, wo N das kleinste n mit $x_n + \Delta x/2 > 9 + l$ ist; dann wird der zur Darstellung der kontinuierlichen Rohrlänge nach 3.3 benötigte Parameter $\alpha = (x_N + \Delta x/2 - 9 - l)/\Delta x$ gesetzt. Da der Anfang des Lippenrohres, $x = 9$, bezüglich des Segmentrasters (x_n) ebenfalls kontinuierlich variiert, muß dort zwischen dem Querschnittssprung von S_M auf S stetig interpoliert werden. Die Interpolation wurde linear bezüglich log S gewählt:

$$\log S_K := \log S_M + \frac{x_K + \Delta x/2 - 9}{\Delta x} \log \frac{S}{S_M},$$

$$K := \begin{cases} M \\ M+1, \end{cases} \quad x_M + \Delta x/2 \gtrless 9,$$

$$S_{K+1} = \ldots = S_N := S.$$

Mit all diesen Annahmen sind die Querschnitts- und Übertragungsfunktion stetig von den Modellparametern abhängig. Die Abhängigkeit ist nicht stetig differenzierbar, wie es eigentlich sein müßte; da jedoch die partiellen Ableitungen als ziemlich grobe Differenzenquotienten bestimmt werden, werden die Unstetigkeiten der Ableitungen "verschmiert".

5.3.4. Anpassung von Spektren (synthetische Vokale)

5.3.4.1. mit Kosinusreihen-Modell

Das Grundschema der spektralen Anpassung wurde schon in 4.3.1. erwähnt: Abtastwerte eines von einem Parameter-Modell des Stimmkanals gegebenen Spektrums werden durch iterative Veränderung der Parameter an ein vorgegebenes Spektrum angepaßt. Die Spektren sind logarithmisch definiert, damit auch Formanten kleiner Intensität berücksichtigt werden. Das variierte Modellspektrum gewinnt man folgendermaßen aus der Modell-Übertragungsfunktion (3:5). Die Koeffizienten des Nennerpolynoms werden mit Nullen auf je 256 Stück ergänzt und dann mittels eines speziellen FFT-Algorithmus für reelle Funktionen [68, 69] fouriertransformiert. Hieraus und aus dem direkt ausgewerteten Zählerpolynom erhält man Abtastwerte des logarith-

mischen Amplitudenspektrums im Abstand von $f_s/256 = 137{,}9$ Hz, entsprechend einem realistischen Spektrallinienabstand (= Grundfrequenz) für männliche Sprecher. Als einfachster Test wurden auch die vorgegebenen Spektren direkt aus den Übertragungsfunktionen von Fant-Querschnittsfunktionen berechnet. Der Frequenzbereich von 3 Formanten (780/N Abtastwerte) wurde berücksichtigt; die Segmentzahl N ist 34 für /i, e, a/, 38 für /o/, 40 für /u/, 39 für /ɨ/. Für die Iteration wurde das einfache Fouriermodell (5:14) mit k = 7 und der Einschränkung $0 \leq b_0 \leq 3$ zugrunde gelegt und in (5:12) $\varepsilon = 10^{-5}$ (anfangs auch $5 \cdot 10^{-5}$) und $\tau = 10$ gesetzt. Die relative Variation der b_j zur Berechnung der partiellen Ableitungen nach (5:13) war D = 1/256. Mindestens 17 Iterationsschritte wurden benötigt, oft die höchste zugelassene Zahl 30 erreicht. Das Ergebnis war i.a. schlecht, außer für /i,e,a/. λ in (5:11) behielt große Werte (10^{-3}...1), was auf eine ungünstige Situation hindeutet. Ein naheliegender Grund ist, daß die geradzahligen Fourierkoeffizienten bei Start vom homogenen Rohr (5 cm^2) nur Effekte 2. Ordnung auf die Formanten haben und sich b_0 sogar fast nur auf die Bandbreiten auswirkt. Als Ausweg aus dieser Schwierigkeit wurden nun versuchsweise während der ersten 5 Iterationsschritte b_0, b_2, b_4, b_6 auf 1,6095 (= log 5), 0, 0, 0 festgehalten, d.h. die Bedingung (5:6) erfüllt; tatsächlich wurden jetzt λ und Φ schnell klein. Bei den bis zu 25 folgenden Schritten nahm der Fehler Φ i.a. weiter stark ab, bei /i, e, o/ um Faktoren 1300, 40000, 8300; λ jedoch blieb groß. Die Ergebnisse sind bei den genannten drei Vokalen ausgezeichnet. Bei /a/ liegt ein gemeinsamer Faktor vor (b_0 falsch), /ɨ/ und besonders /u/ liefern etwas weniger zufriedenstellende Ergebnisse; Φ bleibt in diesen Fällen um drei bis vier Zehnerpotenzen größer. Auf Abbildungen werde verzichtet; denn diese Vorversuche waren noch weit von jeder Praxis entfernt. Dann ist ja das Spektrum mit dem unbekannten Glottisspektrum gewichtet, es ist nur bis auf eine additive Konstante definiert, und seine energiearmen Stellen sind evtl. mit Rauschen aufgefüllt. Um vom Glottisspektrum loszukommen, wurde versucht, von der Übertragungsfunktion H_p auszugehen (2:2), diese ist außerdem unabhängig von Wandeffekten am Rohreingang. Den Eingangsdruck könnte man mit einem schallharten Mikrophon durch die Haut oberhalb des Kehlkopfes messen. Leider waren die Ergebnisse auch für den Idealfall

völlig unbrauchbar, was naheliegt, da die Polfrequenzen von H_p in erster Näherung gerade b_2, b_4, b_6 bestimmen und nicht die für übliche Stimmkanalformen viel wichtigeren b_1, b_3, b_5. Die Verwendung von beiden Polsätzen, d.h. der Eingangsimpedanz Z_1, lieferte überraschenderweise ebenfalls keine sinnvollen Ergebnisse bei Anpassung von $\log|Z_1|$. Vielleicht ist hier die Abhängigkeit des Anpaßfehlers von den Parametern zu kompliziert, so daß das Iterationsverfahren versagt oder in Nebenminima hängenbleibt.

Um einen realistischeren Fall zu betrachten, wurden nun Spektren zugrunde gelegt, die einer Anregung des Stimmkanals (Fant-Vokale) mit den Glottispulsen (4:1) und einem Analyseintervall von einer Grundperiode (256 Abtastwerte) entsprachen. Der Abfall des Spektrums wurde durch 6 dB/Okt. Höhenanhebung ausgeglichen. Die zum Parameter-Modell gehörigen Spektren entsprachen dagegen wieder der Übertragungsfunktion. Da jetzt die absolute Größe aber nicht mehr bekannt war, wurde $\log|H_u|$ durchweg auf den gleichen Mittelwert wie das gegebene Spektrum normiert. Wieder wurden b_0, b_2, b_4, b_6 während der ersten 5 oder auch 12 Iterationsschritte festgehalten. Die Ergebnisse sind i.a. sehr schlecht, außer für /i/; auch wenn 40 statt 30 Schritte zugelassen werden (Fig. 22). Eine Systematik der Fehler ist nicht zu erkennen. Die geraden Fourierkoeffizienten weichen stärker ab, wie zu erwarten; aber auch die Zwischenergebnisse nach den Schritten mit festen b_0, b_2,... stimmen schon nicht mehr mit denen bei Verwendung idealer Übertragungsfunktionenspektren überein. Die Anpassung der tiefen Formanten ist oft schlechter als der hohen, wohl weil sie sich schwächer mit Rohrdeformierungen ändern, vgl. (5:2). Da die Formantbandbreiten meist kleiner als der Spektrallinienabstand sind, ist die von ihnen gelieferte Information sehr anfällig gegen Störungen, z.B. durch das Glottisspektrum.

5.3.4.2. Mit Artikulatormodell

Um die Zahl der Unbekannten zu reduzieren und so evtl. störungsunempfindlicher zu werden, wurde nun das Fouriermodell durch das Artikulatormodell ersetzt. Die Parameter waren mit den Bezeichnungen aus 5.3.3.2: $b_1 = \alpha_J$ (Kieferwinkel); $b_2 = x_p$, $b_3 = d_p$ (Zungenkörper); $b_4 = r$ (Lippenrundung); $b_5 = s$ (Ska-

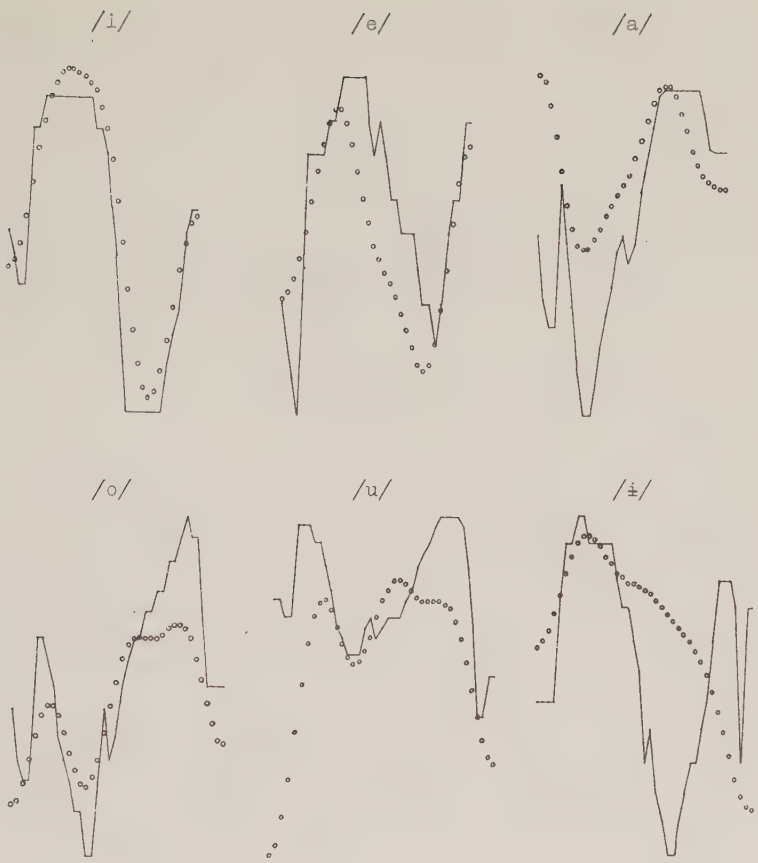


Fig. 22: Querschnittsfunktionen von Fant-Vokalen, ermittelt durch Anpassung der logarithmischen Spektren mit dem Kosinusreihen-Modell.
Ordinatenmaßstab logarithmisch, für jeden Vokal anders.
Original durchgezogen, Ergebnisse Kreise.

lenfaktor) mit den zugelassenen Wertebereichen [0, 0,18], [0, 6], [0,7741935 - 0,116129b_2, 3], ≥ -1 (nach oben begrenzt durch "Mundfläche ≥ 0,01 cm^2"), [0,95, 1,3]. Die obere Grenze 3 für b_3 widerspricht noch (5:16); diese schärfere Bedingung wurde erst später (5.3.5) eingeführt. Alle fünf Parameter gleichzeitig zu bestimmen, liefert keine brauchbaren Ergebnisse. Daher wurden (a) s oder (b) α_J oder (c) s und α_J festgehalten. Anfangswerte waren α_J = 0,05, x_p = 3,5, d_p = 0,7, r = 0, s = 1,12, ungefähr einer neutralen Rohrform entsprechend. Wenn α_J festgehalten wurde, waren seine Werte für /i, e, a, o, u, ɨ/: 35, 83, 90, 146, 76, $42 \cdot 10^{-3}$ (geschätzt nach den Sagittalschnitten in [3], Fig. 2.3-1). In den Fällen (b) und (c) kam für /a/ und besonders für /o/ ein gutes Ergebnis heraus, in (a) nur für /a/, bei /o/ fing sich hier α_J auf dem Randwert, der hier noch 0,3 statt 0,18 betrug. Die Konvergenz ist schneller als für das Kosinusreihen-Modell, z.B. 6 Schritte für /o/ in (c), 8 in (b). /i, e, ɨ/ liefern nie sinnvolle Ergebnisse, die Iteration entfernt sich wenig von einer neutralen Rohrform, abgesehen von Änderungen der Lippenform; dennoch wird das Verfahren oft schnell als "konvergiert" betrachtet, und für /ɨ/ stimmt sogar das Spektrum recht gut. /u/ ergibt qualitativ richtige Lösungen nach (b) und (c), jedoch mit zu starken Einschnürungen; nach (a) kommt eine sehr weite Mundöffnung und eine Einschnürung im Rachen heraus. Dies liegt offenbar wieder daran, daß für α_J in (a) noch ein Wertebereich bis 0,3 zugelassen war. Die Spektrumsanpassung ist in allen drei Fällen recht gut; bei (a) liegen die ersten beiden Formanten etwas zu hoch. Das Verhalten des Marquardt-Parameters λ ist unsystematisch; meist bleibt er groß. Fig. 23 zeigt Beispiele: (b)/a, o, u, ɨ/. Oben sind die Querschnittsfunktionen logarithmisch aufgetragen (gestrichelt: vorgegebene Form; ausgezogen: Iterationsergebnis; Parallelverschiebung ist nicht schlimm zu bewerten), die Abszisse läuft von unten nach oben, die Ordinate von rechts nach links. Darunter ist ein Sagittalschnitt des Stimmkanals abgebildet, rechts unten die Mundform, die Strecke darüber gibt die Lippenrohrlänge an.

Bei Verwendung des Artikulatormodells ist deutlich zu beobachten, daß sehr unterschiedliche Rohrformen ähnliche Spektren liefern können (vgl. (a)/u/). Eine hinreichende Begrenzung der erlaubten Parameterbereiche ist notwendig, um ein

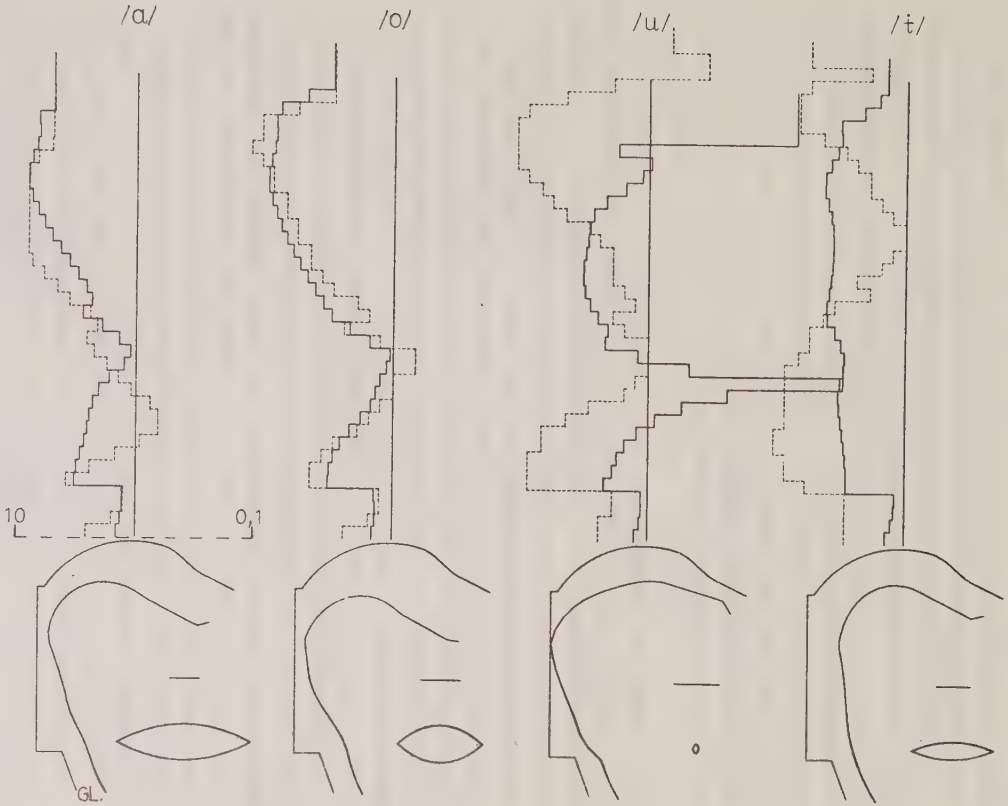


Fig. 23: Querschnittsfunktionen von Fant-Vokalen, ermittelt durch Anpassung der logarithmischen Spektren mit dem Artikulatormodell; darunter zugehörige Sagittalschnitte und Mundöffnungs-Formen. Im übrigen siehe Text

Weglaufen der Parameter bei der Iteration zu verhindern (vgl. (a)/o, u/); jedoch bergen solche Begrenzungen die Gefahr in sich, daß sich die Iteration in einem Randminimum fängt. Dies geschah bei (a) außer bei /e/ immer für irgendeinen Parameter, bei (b)/e/ für s = 1,3, bei (c)/e/ für r = -1, (c)/u/ für x_p = 0. Wie aufgrund der teilweisen Kompensierbarkeit von Kieferbewegungen durch Zungenbewegungen zu vermuten war, verbessert ein bekanntes, festgehaltenes α_J die Ergebnisse beträchtlich. Allgemein läßt sich jedoch sagen: Die Anpassung von logarithmischen Spektren liefert nicht mit der nötigen Zuverlässigkeit brauchbare Ergebnisse.

5.3.5. Prädiktive Anpassung

5.3.5.1. Die beiden Verfahren

Schon bei der Untersuchung der Verfahren von Atal und Wakita (5.2) wurde erwähnt, daß man unabhängig von Annahmen über das Glottissignal und von Formantverschmierungseffekten werden könnte, wenn man nur Signalabschnitte analysiert, für die das Eingangssignal verschwindet. Dann hängt das Verhalten des Ausgangssignals nur von den Rohreigenschaften ab, außerdem entspricht die Modellannahme einer schallharten Glottis dann gut der Wirklichkeit. Die prädiktive Analyse nach dem Autokovarianzverfahren (A.2), die sich hier anbietet, hat den Nachteil, daß die geringe Länge des Analyseintervalls Ungenauigkeiten bewirkt und Pole außerhalb des Einheitskreises vorkommen können. Es wird also keine großen Vorteile bringen, aus einem derartigen Prädiktor abgeleitete Größen als die anzupassenden akustischen Größen zu wählen und etwa die Spektren für Anpassung nach 5.3.4 auf diese Weise zu ermitteln. Nun stellt aber Prädiktion selbst schon einen Anpaßvorgang im Sinne einer kleinsten Fehlerquadratsumme (A:4) dar. Es liegt also nahe, vorhergesagte Abtastwerte des Signals während des Glottisverschlußintervalls gegen die wirklichen Werte (oder das inversgefilterte Signal gegen Null) anzupassen. Dies läuft auf eine Prädiktion oder Inversfilterung unter Nebenbedingungen hinaus: Man bilde den besten Prädiktor, der mit den möglichen Übertragungsfunktionen des benutzten Parameter-Modells verträglich ist. Jedem Parametervektor $\underline{b}$ muß also ein Prädiktor $A(\underline{b};z)$ zugeordnet werden, der aus der Übertragungsfunktion berechnet

wird. Beschränkt man den Begriff "Prädiktor" hier ausnahmsweise nicht auf Polynome, so liefert z.B. die inverse Übertragungsfunktion H_u^{-1}, evtl. geeignet normiert, einen solchen. Auch das Nennerpolynom von H_u ist ein Prädiktor; ebenso (Nenner von H_u)/(beliebiges Polynom), wo letzteres den Frequenzgang geeignet beeinflußt. Ein Prädiktor mit Nenner bedeutet einfach, daß der vom Zählerpolynom erzeugte Prädiktionsfehler noch einmal rekursiv gefiltert wird. Dabei müssen Anfangswerte bzw. Integrationskonstanten vorgegeben werden, weil die "Vergangenheit" des rekursiv gefilterten Signals nicht bekannt ist.

Ein anderes Verfahren, einen Prädiktor aus der Übertragungsfunktion herzuleiten, ist folgendes. Man berechnet zunächst das Anfangsstück der AKF der Impulsantwort; für (Nenner von H_u)$^{-1}$ geht das nach (A:13), (A:10), für den Zähler direkt, beide werden gefaltet. Aus der AKF erhält man dann nach (A:11), (A:12) das Prädiktorpolynom. (Dieses Verfahren wurde bereits in 5.3.2 benutzt, ohne dort genauer beschrieben worden zu sein.)

Die Modell-Übertragungsfunktion wird für d = 0,5 cm Segmentlänge, also f_s = 35,3 kHz berechnet. Die untersuchten Vokale sind ebenfalls mit dieser Frequenz abgetastet, jedoch auf 4,4 kHz bandbegrenzt. Es sollte also genügen, nur jeden vierten Wert für die Anpassung zu verwenden, was effektiv auf eine f_s-Vierteilung hinausläuft ($f_s/4$ = 8825 Hz). Bei Benutzung der ersten Methode zur Prädiktorbestimmung werden natürlich auch die dazwischen liegenden Werte bei der Inversfilterung gebraucht, da ja der Nenner der Modell-Übertragungsfunktion unverändert verwendet wird. Bei der zweiten Methode dagegen kann man die AKF der Impulsantwort tiefpaßfiltern, nur jeden vierten Wert berücksichtigen und daraus ein zu $f_s/4$ gehöriges Prädiktorpolynom ermitteln, was die Verwendung von wirklich mit 8825 Hz abgetasteter Sprache trotz d = 0,5 cm ermöglicht.

Es ergeben sich also die folgenden beiden Verfahren, das Inversfilter (den Prädiktor) zu konstruieren:

1.) <u>"Nenner-Verfahren":</u> $H_u \mapsto \frac{\text{Nenner von } H_u}{\text{Polynom}}$,

2.) <u>"AKF-Verfahren":</u> $H_u \mapsto$ AKF der Impulsantwort $\mapsto$ tiefpaßgefilterte und neuabgetastete AKF $\mapsto$ Prädiktor bei $f_s/4$.

Als Tiefpaßfilter bei 2.) wurde wieder (4:2) benutzt. Zunächst wurde die AKF einfach mit der Impulsantwort des Tiefpasses gefaltet. Die dann entstehende Funktion (R_i) ist jedoch keine

wirkliche AKF mehr, da ihr Spektrum und die Matrix (R_{i-j}) nicht mehr positiv definit sind. Tatsächlich muß man mit der AKF der Tiefpaß-Impulsantwort falten; hierdurch werden auch höhere Frequenzen, die bei der Neuabtastung Aliasing-Fehler verursachen können [23, 24], besser unterdrückt.

5.3.5.2. Artikulatormodell mit Nennerverfahren (synthetische Vokale)

Zunächst wurde einfach der Nenner der Übertragungsfunktion (3:5) nach Durchdividieren mit dem ersten Koeffizienten als Prädiktorpolynom verwendet. Die Prädiktion begann am Anfang des Glottisverschlußintervalls; für jeden 4. Abtastwert (im ganzen 20 Stück) wurde der vorhergesagte Wert mit Hilfe des Iterationsprogramms (5.3.2) an den wirklichen angepaßt. Die Modellparameter und ihre Anfangswerte waren die gleichen wie schon in 5.3.4.2. Festgehalten wurden α_J und s oder α_J allein (auf den gleichen Werten wie in 5.3.4.2). Die Ergebnisse für die Vokale nach Fant-Querschnittsfunktionen sahen erfolgversprechend aus, zumindest für /a,o,u/; für /i,e,ɨ/ entfernte sich die Iteration wenig von einer neutralen Rohrform. Die Konvergenz war schnell (5 bis 16 Schritte); die wesentlichste Abnahme des Prädiktionsfehlers geschah i.a. schon beim ersten oder den ersten beiden Schritten; der Marquardt-Parameter λ war oft klein. Es fiel auf, daß für /i,e,a/ immer eine viel zu große Lippenöffnung und eine zu große Rohrlänge herauskam (obwohl für /a/ das Ergebnis sonst gut war). Beides könnte zusammenhängen: Durch die große Lippenöffnung wird das Lippenrohr praktisch unwirksam gemacht. Daher wurde bei allen weiteren Versuchen mit festem Skalenfaktor s für /i,e,a/ s = 1 gesetzt, für /o,u,ɨ/ weiterhin s = 1,12. In diesem Unterkapitel sind im folgenden α_J und s durchweg fest.

Da es sein könnte, daß das Artikulatormodell die Fant-Querschnittsfunktionen überhaupt nicht gut darstellen kann, sollten im folgenden auch solche synthetischen Vokale zugrunde gelegt werden, die mit dem Artikulatormodell synthetisiert worden waren. Dazu wurden dessen Parameter so bestimmt, daß statt gewisser akustischer Größen die logarithmischen Querschnittsfunktionen selbst (bis auf eine Konstante) optimal an die Fant-Funktionen angepaßt wurden. Das Ergebnis zeigt Fig. 24. Die

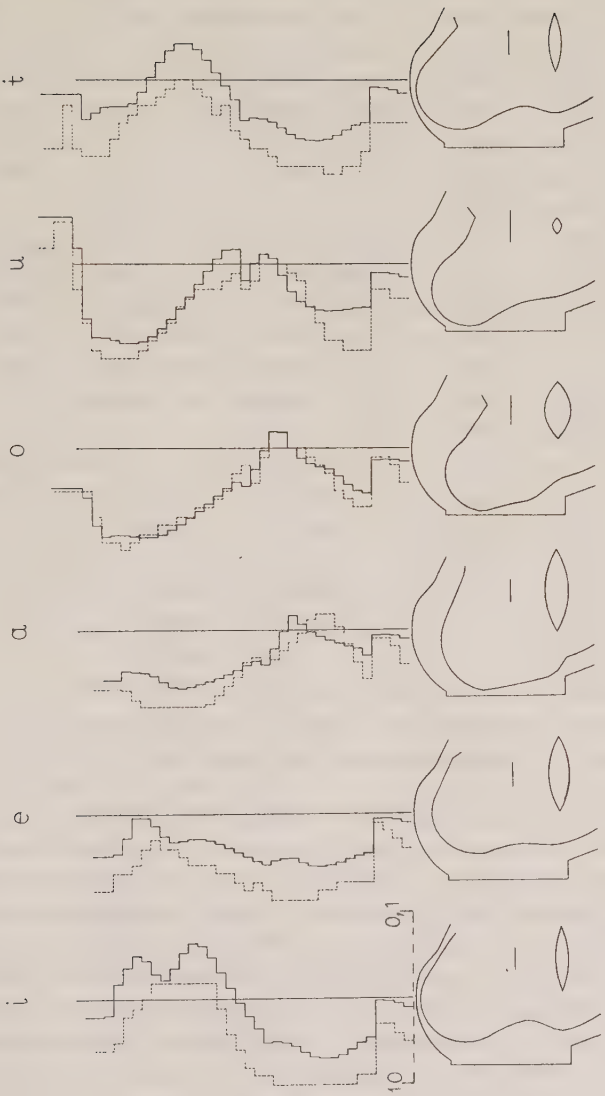


Fig. 24: Direkte Anpassung des Artikulatormodells an logarithmische Querschnittsfunktionen nach Fant (diese gestrichelt)

Übereinstimmung ist recht gut, bis auf eine additive Konstante; die Formanten liegen jedoch etwas anders. Tab. 1 zeigt die Formanten für beide Vokal-Sätze und die Parameter des Artikulatormodells. Der Vokal-Satz nach dem Modell werde mit "FMOD"

		F_1	F_2	F_3	F_4	B_1	B_2	B_3	B_4
Fant	/i/	258	2285	2863	3705	47	17	251	74
	/e/	438	1958	2739	3389	23	54	208	344
	/a/	718	1131	2526	3638	44	36	53	153
	/o/	543	860	2369	3494	39	23	16	15
	/u/	261	612	2384	3744	48	14	1	4
	/ɨ/	309	1432	2339	3374	36	71	11	39
Fmod	/i/	318	2202	2686	3812	85	25	29	45
	/e/	485	1800	2780	3688	38	25	82	227
	/a/	727	1320	2748	3783	43	33	89	130
	/o/	558	929	2458	3478	33	28	21	27
	/u/	286	706	2691	3368	51	24	4	5
	/ɨ/	346	1356	2368	3514	67	26	8	22

		$10^3\alpha_J$	x_p	d_p	$10^4 r$	s
Fmod	/i/	35	5,039	1,691	984	1,00
	/e/	83	5,781	1,385	109	1,00
	/a/	90	0,704	0,692	-737	1,00
	/o/	146	3,337	0,778	2305	1,12
	/u/	76	2,802	1,418	6227	1,12
	/ɨ/	42	4,377	1,652	1538	1,12

Tab. 1: Formanten für FANT- und FMOD-Vokale, Modellparameter für FMOD-Vokale

bezeichnet. Die so ermittelten Modellparameter wurden nun als Startwerte für die prädiktive Anpassung an Fant-Vokale benutzt, um zu sehen, ob bei Iterationsbeginn nahe der Lösung das Ergebnis sicher herauskommt. Tatsächlich waren alle Ergebnisse gut und wurden schnell erreicht (2-11 Schritte), lediglich bei /u/ lief die Iteration in eine falsche Richtung (am Velum lag eine Einschnürung, der Mund war weit geöffnet) und fing sich durch die Lippenrundungs-Begrenzung $r \geq -1$ in einem Randminimum. Dies zeigt wieder die Gefährlichkeit solcher willkürlichen Begrenzungen, die leider nötig sind.

Als nächstes wurde nun die Anpassung an beide Vokalsätze (FANT und FMOD) versucht. Da bei der geringen Modellparameter-

zahl die ersten drei bis vier Formanten ausreichen müßten und die Komponenten dicht unter 4,4 kHz höchstens zusätzliche Fehlerquellen darstellen, wurde das synthetische Signal im folgenden noch einmal tiefpaßgefiltert. Die Impulsantwort entsprach (4:2), die "13" war jedoch durch "9" ersetzt, so daß das Spektrum schon bei 3 kHz abfiel und bei 4,4125 kHz Null war (Fig. 9, gestrichelt). Die Startwerte für die Iteration wurden von jetzt an als $x_p = 3{,}5$, $d_p = 1{,}2$, $r = 0{,}1$ angesetzt; wie die FMOD-Parameter in Tab. 1 zeigen, waren die früheren Werte $d_p = 0{,}7$, $r = 0$ i.a. zu klein. Hierauf könnte das Versagen der Anpassung für /i,e,ɨ/ beruht haben, das auch schon bei der Spektrumsanpassung zu beobachten war. Als zusätzliche Kontrolle wurden aus den Nullstellen des Nennerpolynoms der Übertragungsfunktion die Formanten ermittelt, sie konnten mit den Werten in Tab. 1 verglichen werden.

Die Ergebnisse waren nach diesen Änderungen z.T. besser, z.T. schlechter als vorher. Bei /ɨ/ und /e/ war das Rohr noch immer weitgehend neutral, bei Fmod-/u/ trat ein Verschluß im Rachen auf. Nach wie vor waren die Lippenquerschnitte bei /i, e,a,ɨ/ viel zu groß. Die Formantfrequenzen wichen teils bis zu 200 Hz von den richtigen Werten ab. Auffällig war der besonders große Fehler des 1. Formanten, zumal wenn dieser tief lag. Als Grund wurde folgendes vermutet: Der Zähler von H_u in (3:5) hat eine Nullstelle bei $w_2/w_1 \approx 1$, der Nenner ebenfalls in der Nähe dieses Wertes, die eine starke Absenkung der tiefen Frequenzen bei der Inversfilterung mit dem Nenner allein bewirkt. Diese gehen in das zu minimierende Fehlerquadrat dann nur schwach ein. Ein Ausgleich durch rechnerische Tiefenanhebung von kurzen Abschnitten des Signals selbst ist nicht ohne weiteres möglich, da hierbei ein nur langsam abklingender Einschwingvorgang entsteht. Stattdessen wurde der Prädiktionsfehler tiefenangehoben durch Verwendung einer Inversfilterfunktion mit Nenner. Zunächst wurde als Nenner $1 - (w_2/w_1)z^{-1}$ gewählt, dann einfach $1 - z^{-1}$, was besser ging. Dies entspricht einer unbestimmten Integration (genauer: Summation) des vom Zähler allein erzeugten Prädiktionsfehlers; der Mittelwert des Ergebnisses wurde auf Null normiert und dann diese Funktion gegen Null angepaßt. Jetzt waren die Ergebnisse oft sehr gut. Bei /a,o,ɨ/ wurde die Iteration allerdings meist schon beim ersten Schritt als konvergiert betrachtet, da schon die An-

fangswerte nahe am Ergebnis lagen (der x_p-Anfangswert war jetzt 4). Verzichtete man auf die zusätzliche Tiefpaßfilterung bei 3 kHz, so wurden wieder einige Schritte ausgeführt. Ob das Ergebnis gut war oder nicht, hing leider ohne erkennbare Systematik von vielen Kleinigkeiten ab: der Wahl der Anfangswerte, der oberen Grenzfrequenz, der Genauigkeitsschranke ε in (5:12), der b_j-Variation D in (5:13) (was evtl. auf einen Einfluß der endlichen Rechengenauigkeit hindeutet). Ein D von 1/64 erwies sich oft als günstiger als 1/256. Einige Ergebnisse sind in Fig. 25 a,b dargestellt, $\varepsilon = 10^{-5}$, D = 1/64, (a) zusätzlich tiefpaßgefiltert, (b) nicht. Fig. 26 zeigt zu (a) gehörige Verläufe der relativen Prädiktionsfehlerquadratsumme (Achtung: verschiedene Ordinatenmaßstäbe!) und der Modellparameter im Zuge der Iteration. Die Konvergenz ist i.a. schnell. Tab. 2 zeigt Formantwerte der Ergebnisse (vgl. mit Tab. 1).

	a: F_1	F_2	F_3	F_4,	b: F_1	F_2	F_3	F_4
Fant /i/	288	2252	2776	3844	525	1849	2722	3868
/e/	446	1962	2771	3844	596	1305	2646	3681
/a/	607	1287	2651	3681	653	1097	2734	3842
/o/	588	985	2394	3329	552	923	2429	3366
/u/	197	1213	1697	2456	293	824	2551	3321
/ɨ/	457	1228	2330	3361	358	1472	2342	3474
Fmod /i/	326	2202	2690	3815	299	2203	2688	3807
/e/	526	1804	2698	3817	503	1811	2694	3812
/a/	607	1287	2651	3681	607	1287	2651	3681
/o/	588	985	2394	3329	588	985	2394	3329
/u/	285	708	2721	3375	510	1142	2343	3330
/ɨ/	457	1228	2330	3361	457	1228	2330	3361

Tab. 2: Formantfrequenzen der Ergebnisse nach Fig. 25

Zusammenfassend kann gesagt werden: Die Ergebnisse der Iteration mit dem Artikulatormodell nach dem "Nennerverfahren" mit einfach integriertem Prädiktionsfehler können sehr gut sein; ob sie es wirklich sind, hängt aber in undurchsichtiger, für jeden Vokal anderer Weise von den genauen Umständen ab. Die Anpassung an die FMOD-Vokale liefert keine sehr viel besseren Ergebnisse als die an die FANT-Vokale. Die durchschnittliche Rechenzeit beträgt weniger als eine Minute pro Vokal.

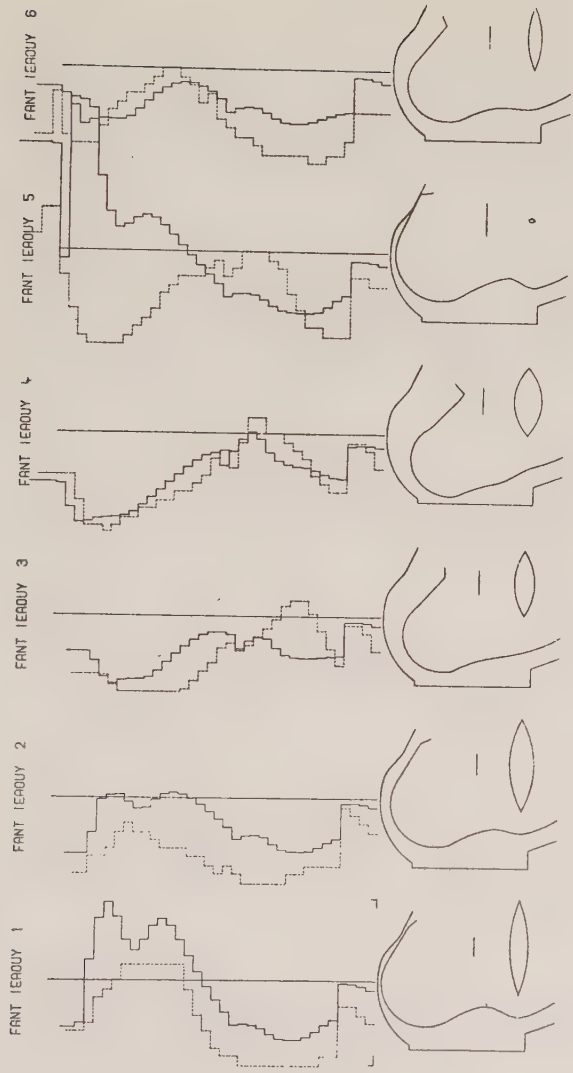


Fig. 25: Prädiktive Anpassung des Artikulatormodells an Fant-Vokale nach dem Nennerverfahren (siehe Text)
(a) Grenzfrequenz 3 kHz

Forts.

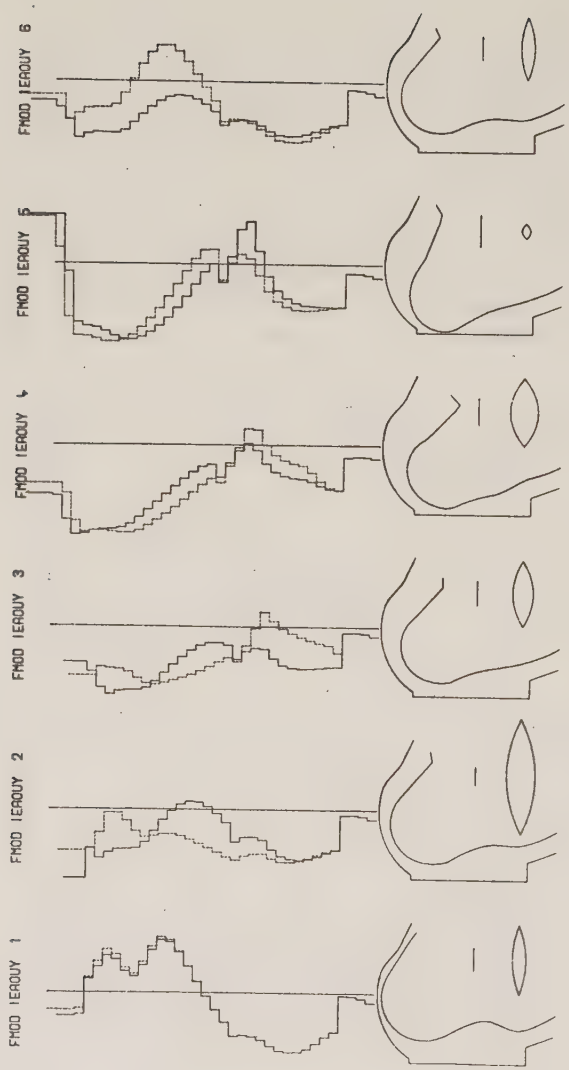


Fig. 25 a (Fortsetzung)

Forts.

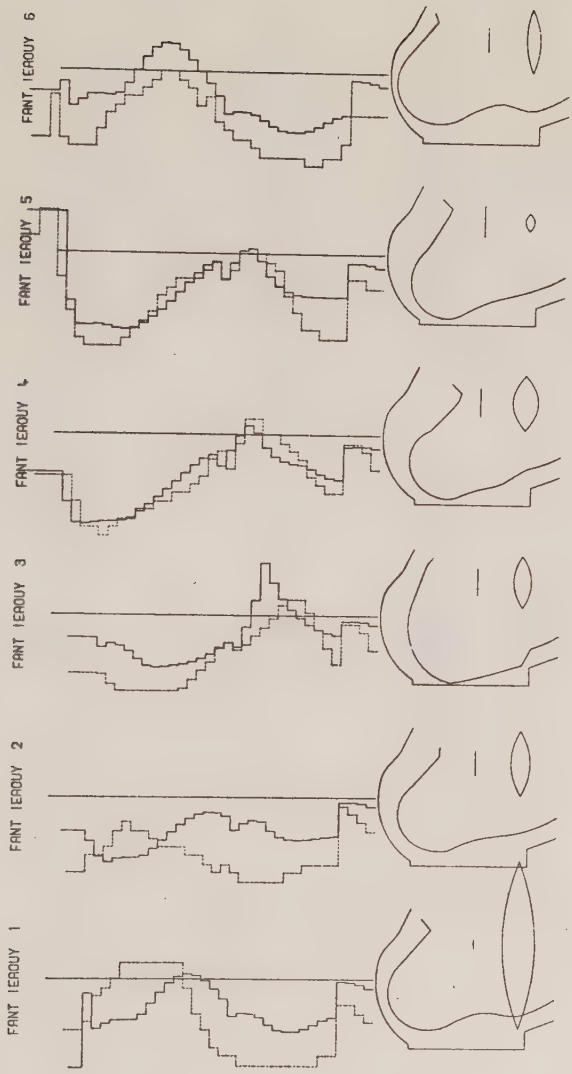


Fig. 25 b: Grenzfrequenz 4,4 kHz

Forts.

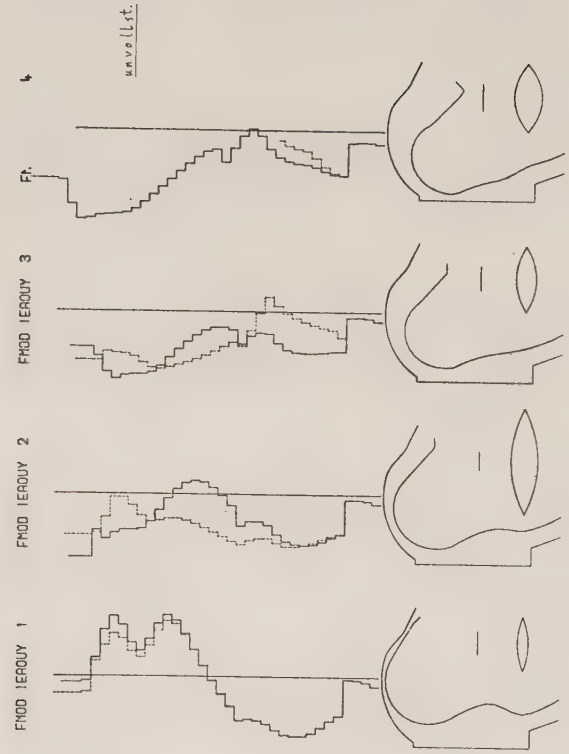


Fig. 25 b (Fortsetzung)

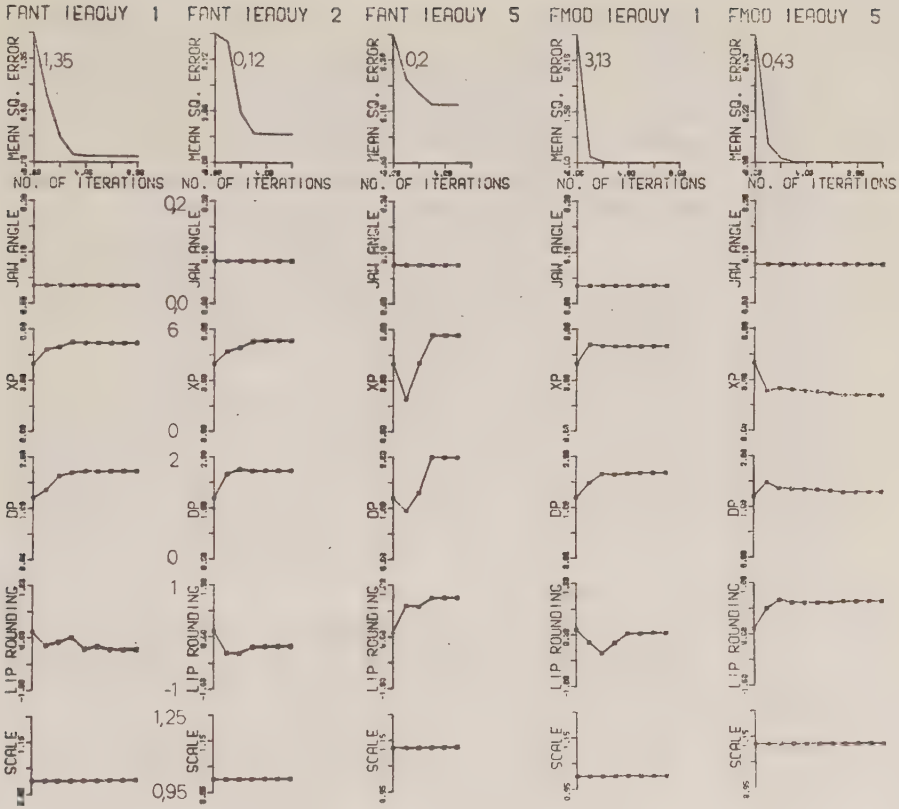


Fig. 26: Verlauf der relativen Fehlerquadratsumme (verschiedene Ordinatenmaßstäbe!) und der Modellparameter α_J (fest), x_p, d_p, r, s (fest) während der Iteration für einige Vokale nach Fig. 25 a

5.3.5.3. Artikulatormodell mit AKF-Verfahren
(synthetische Vokale)

Das "AKF-Verfahren" erlaubt es, mit einem einfachen Prädiktorpolynom zu arbeiten, das aufgrund der Abtastfrequenz-Vierteilung nur 1/4 bis 1/3 so lang ist wie das beim "Nennerverfahren" benutzte Nennerpolynom von H_u. Als Prädiktorgrad wurde 12 gewählt; wurde bei der Lösung der Gleichungen (A:11), (A:12) durch Ungenauigkeiten ein $|a_n^n| \geq 1$, so durfte der Grad kleiner sein, aber nicht kleiner als 8 (dann wurde die Rechnung für diesen Vokal abgebrochen). Eine besondere Neigung zum Versagen aus diesem Grund ergibt sich bei /u/, dessen 3. und 4. Formant (im Modell) sehr kleine Bandbreiten haben und so leicht zu Polen außerhalb des Einheitskreises führen. Aber auch bei /i,e,ɨ/ versagte das Verfahren oft.

Die Ergebnisse der Versuche hingen weniger als erwartet davon ab, ob die Tiefpaßfilterung der AKF (vgl. 5.3.5.1) mit der Impulsantwort (4:2) oder - theoretisch richtig - mit deren AKF erfolgte. Dies lag wohl daran, daß einmal die nicht-positiven Spektralkomponenten der Impulsantwort sehr klein sind, zum anderen die AKF durch den höheren Rechenaufwand zu größeren Ungenauigkeiten führt.

Die Zahl der benötigten Iterationsschritte war ähnlich wie nach dem Nennerverfahren (typisch 5 bis 13). Allgemein mußten wieder s allein oder s und α_J festgehalten werden. War nur s fest, kamen häufiger Versager wegen $|a_n^n| \geq 1$ vor, es wurden mehr Schritte benötigt, und die Ergebnisse waren meist etwas schlechter, in Einzelfällen aber auch besser. Ob das Signal zusätzlich bei 3 kHz tiefpaßgefiltert war oder nicht, führte in unsystematischer Weise teils zu Verbesserungen, teils zu Verschlechterungen. Fig. 27 und Fig. 28 zeigen Beispiele: Die Sprache war bei 3 kHz bandbegrenzt, die Filterung der AKF erfolgte nur in Fig. 28 auf die theoretisch richtige Weise, die Anfangswerte waren in Fig. 27 $x_p = 3{,}5$, $d_p = 1{,}2$, $r = 0{,}1$, in Fig. 28 war $x_p = 4 + 10\alpha_J$; s und α_J wurden auf den gleichen Werten wie in 5.3.5.2 festgehalten. Die $\underline{b}$-Variation D war in Fig. 27 1/256, in Fig. 28 1/64. Tab. 3 gibt die Formantwerte der Lösungen wieder

Auch das Verhalten bei Wahl der Anfangswerte nahe der Lösung wurde wieder untersucht; die Anfangswerte waren die FMOD-

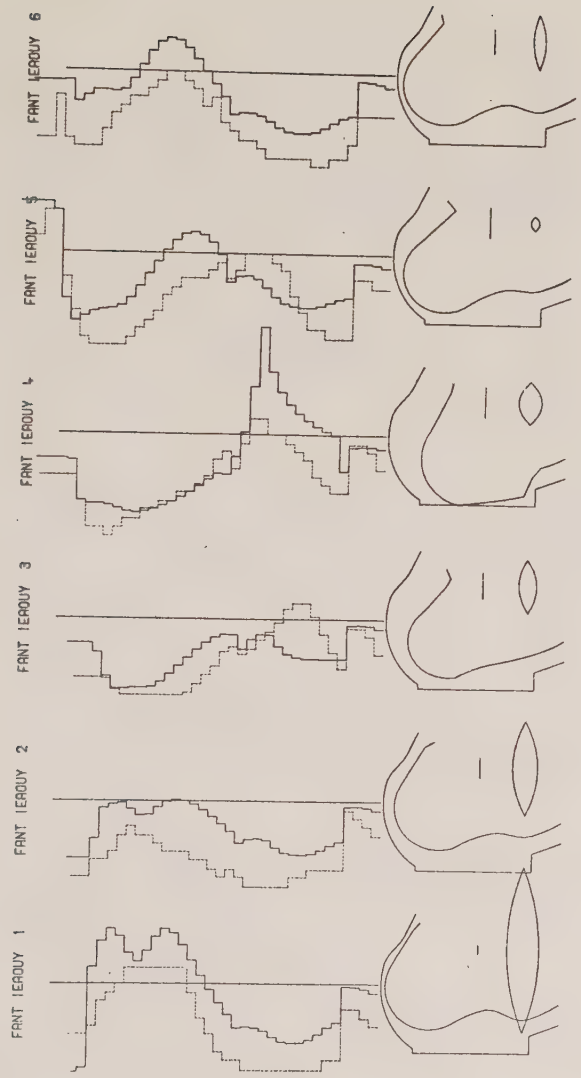


Fig. 27: Prädiktive Anpassung des Artikulatormodells an Fant-Vokale nach dem AKF-Verfahren (siehe Text), falsche Tiefpaßfilterung der AKF

Fonts.

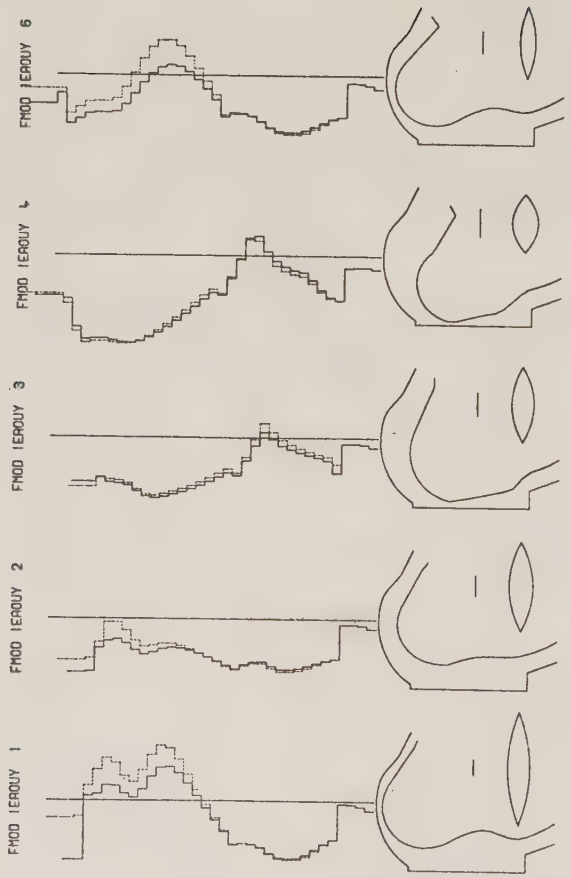


Fig. 27 (Fortsetzung)

Fig. 28: Prädiktive Anpassung des Artikulatormodells an Fant-Vokale nach dem AKF-Verfahren (siehe Text), richtige Tiefpaßfilterung der AKF

Forts.

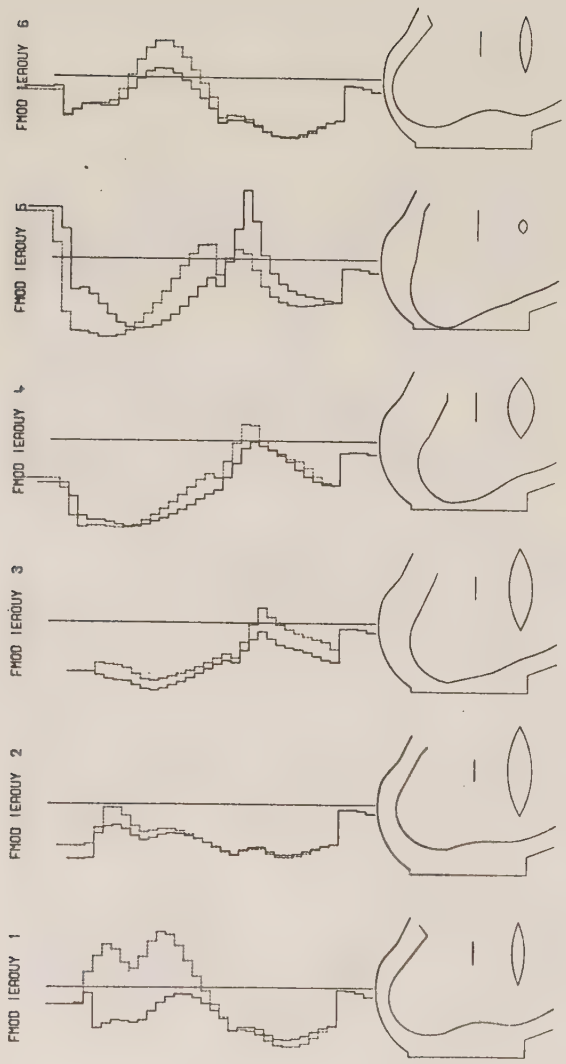


Fig. 28 (Fortsetzung)

		Fig. 27			Fig. 28		
		F_1	F_2	F_3	F_1	F_2	F_3
Fant	/i/	314	2290	2834	301	2276	2821
	/e/	461	1943	2764	462	1945	2771
	/a/	585	1166	2700	540	1020	2664
	/o/	555	732	2364	548	795	2368
	/u/	293	693	2451	hat versagt		
	/ɨ/	341	1427	2337	339	1426	2336
Fmod	/i/	373	2204	2760	482	1525	2586
	/e/	539	1783	2750	541	1784	2752
	/a/	697	1326	2732	690	1356	2742
	/o/	568	930	2466	563	1042	2462
	/u/	hat versagt			299	697	2719
	/ɨ/	408	1373	2345	381	1368	2301

Tab. 3: Formantfrequenzen der Ergebnisse nach Fig. 27, 28

Parameter. Das Ergebnis für die Fant-Vokale zeigt Fig. 29, es ist von ähnlicher Güte wie für das Nennerverfahren. Für die Fmod-Vokale sollten schon die Startwerte das exakte Ergebnis liefern; tatsächlich entfernt sich die Iteration aber in 3 bis 9 Schritten ein wenig von diesen, wohl durch Rechenungenauigkeiten. Die Abweichungen sind klein, auf eine Abbildung werde verzichtet. Nur bei /i/ zeigt sich - wie übrigens meistens - die Tendenz zu einer viel zu großen Lippenöffnung.

Zusammenfassend besteht in der Güte der Ergebnisse kein großer Unterschied zum Nennerverfahren. Die Tendenz, bei einer neutralen Rohrform stehenzubleiben, fehlt beim AKF-Verfahren; dafür kommt es öfters zum Versagen durch Pole außerhalb des Einheitskreises bei der Prädiktorberechnung. Die Abhängigkeit der Ergebnisse von den genauen Umständen ist ähnlich unsystematisch wie in 5.3.5.2, aber vielleicht etwas geringer. Leider führen die zusätzlichen Rechenschritte zu einer Steigerung der Rechenzeit auf mindestens das Doppelte (zwei Minuten pro Vokal), und der Kernspeicherbedarf kann auf der H632 mit 24 K Speicherworten nur noch mit programmiertechnischen Kunstgriffen erfüllt werden.

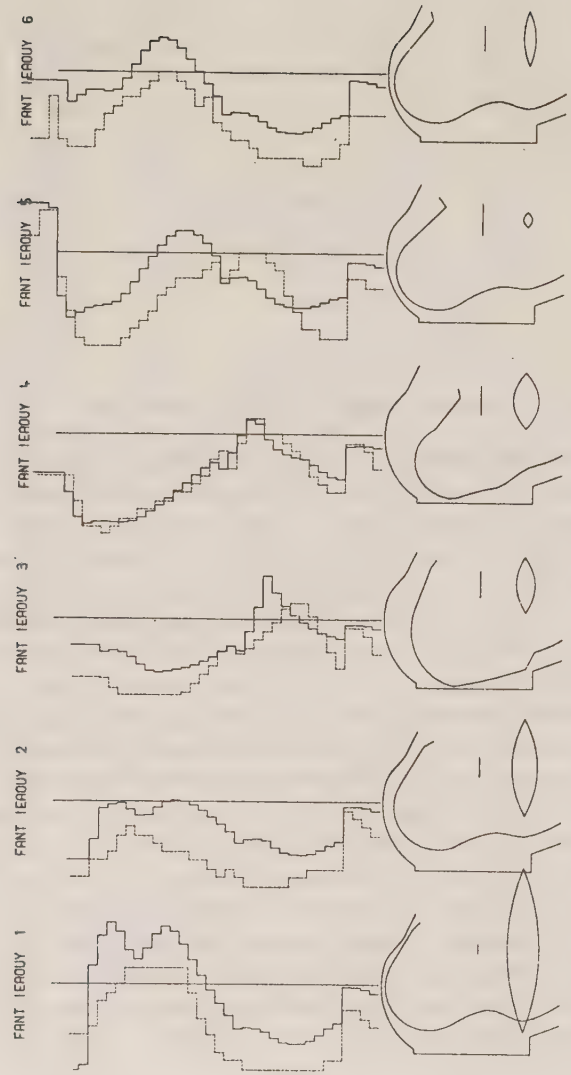


Fig. 29: Prädiktive Anpassung des Artikulatormodells an Fant-Vokale nach dem AKF-Verfahren (siehe Text), Start nahe der Lösung von den FMOD-Parametern (Ergebnis von Fig. 24)

5.3.5.4. Kosinusreihenmodell (synthetische und natürliche Vokale)

Das Artikulatormodell weist zwei Nachteile auf: 1. Es ist äußerst kernspeicher- und rechenintensiv, daher auch anfällig gegen Effekte endlicher Rechengenauigkeit. 2. Der Zusammenhang zwischen Modellparametern und Übertragungsfunktion ist äußerst kompliziert, so daß gleichgroße Variationen der Parameter höchst unterschiedlich starke Wirkungen haben können. Daher sollte auch das Kosinusreihenmodell, jetzt in seiner verbesserten Form (5: 15), nochmals mit den beiden prädiktiven Anpaßverfahren nach 5.3.5.1 verwendet werden, obwohl es die Menge der möglichen Querschnittsfunktionen weit weniger einschränkt. Der Wertebereich des Fourierkoeffizienten b_0 wurde auf [1, 3] beschränkt. Zugrunde gelegt wurden immer die Fant-Vokale mit zusätzlicher Tiefpaßfilterung bei 3 kHz.

Schon eine gleichzeitige iterative Bestimmung der Fourierkoeffizienten $b_0,\ldots,b_6$ mit den Startwerten $b_0 = 1{,}6094$, $b_1 = \ldots = b_6 = 0$ lieferte nach dem Nennerverfahren (mit integriertem Prädiktionsfehler wie in 5.3.5.2) recht gute Ergebnisse, die Formanten stimmten meist bis auf größenordnungsmäßig 10 Hz, die Iterationsschrittzahl war 7 bis 27. Das AKF-Verfahren arbeitete schlechter und versagte bei /o/, die Formanten stimmten ebenfalls nicht so gut. Die Ergebnisse verbesserten sich noch (wie schon in 5.3.4.1), wenn während der ersten 4-5 Schritte b_0, b_2, b_4, b_6 festgehalten wurden; das AKF-Verfahren war aber auch hier schlechter, bei /a/ war der Fehler erheblich, bei /e/ und /ɨ/ versagte das Verfahren. Fig. 30 a,b zeigt die beiden Ergebnis-Sätze. Für das Nennerverfahren wurde auch die Zahl k der Parameter von 6 bis 9 variiert (höchster Koeffizient b_5 bis b_8). Der Unterschied zwischen k = 6 und k = 7 war sehr gering. k = 8 ergab eine kleine aber deutliche Verbesserung; nach den Überlegungen zu (5:1), (5:2) wurde jetzt gleichsam auch der vierte Formant berücksichtigt (Fig. 30 c). k = 9 führte zu anscheinend sinnlosen Welligkeiten der Querschnittsfunktion. Tab. 4 zeigt die Formantwerte und die relative Fehlerquadratsumme für k = 7 und 8. Die Neigung zu einer Parallelverschiebung der Kurven im Sinne eines zu großen b_0 ist in ihren Ursachen nicht geklärt (bei /u/ bzw. /o,u/). Ein Versuch, H_u^{-1} statt (Nenner von H_u)/$(1 - z^{-1})$ als Inversfilter zu benutzen, beseitigte diese Neigung nicht, sondern führte lediglich zu einer Gesamtver-

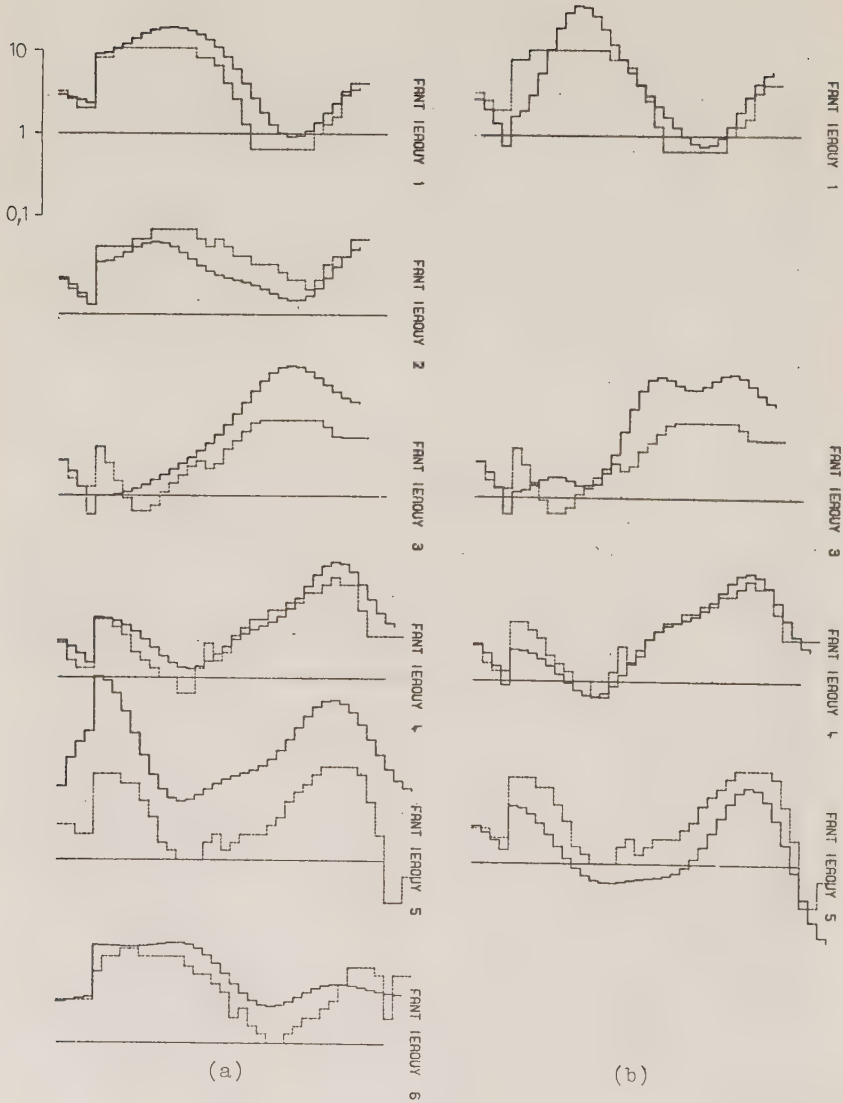


Fig. 30: Prädiktive Anpassung des Kosinusreihenmodells an Fantvokale (siehe Text); (a) Nennerverfahren, (b) AKF-Verfahren, $k = 7$

Forts.

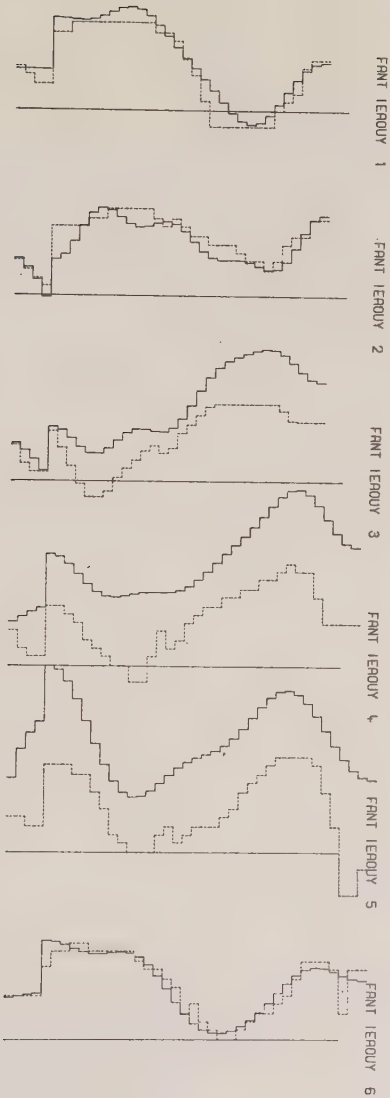


Fig. 30 (c): Wie (a), aber k = 8

k		F_1	F_2	F_3	B_1	B_2	B_3	$10^5\Phi/\Sigma s_i^2$
7	/i/	257	2297	2842	30	24	216	317
	/e/	445	1986	2733	35	57	168	144
	/a/	731	1286	2496	52	47	37	30,6
	/o/	554	881	2377	37	25	12	60,3
	/u/	247	589	2378	11	11	8	41
	/ɨ/	314	1441	2334	24	39	16	210
8	/i/	256	2286	2838	38	26	226	250
	/e/	446	1999	2704	26	49	137	47,2
	/a/	730	1175	2535	55	48	41	34,5
	/o/	539	860	2364	43	42	24	23,5
	/u/	248	592	2380	12	11	11	39,5
	/ɨ/	308	1438	2334	36	75	12	128

Tab. 4: Formanten und relative Fehlerquadratsumme für die Ergebnisse nach Fig. 30 a,c

schlechterung. b_0 ist natürlich allgemein nicht sehr scharf definiert, da seine akustische Wirkung im wesentlichen die Strahlungsdämpfung und den Wandadmittanz-Einfluß betrifft und klein ist.

Da die prädiktive Anpassung des Kosinusreihenmodells die besten Ergebnisse liefert und wenig rechen- und kernspeicherintensiv ist, sollte es auch auf natürliche Vokale angewendet werden. Da auf dem Glottisverschlußintervall gearbeitet werden soll, ist hier zusätzlich eine Routine zum Suchen des Glottisverschlußzeitpunktes nach 4.3 ins Programm einzubauen (beim Artikulatormodell läßt dies der verfügbare Kernspeicher nicht zu). Dieses arbeitet bei $f_s/4 = 8825$ Hz. Die aufgenommenen Vokale sind /i,e,a,o,u,y/ ("TEST DIFF, 2"), gesprochen vom Autor; die Aufnahme erfolgte nach 4.2. Die Parameterzahl war $k = 6$, b_0, b_2, b_4 wurden während der ersten vier Schritte festgehalten. Ein grundsätzlicher Unterschied gegenüber den künstlichen Vokalen trat beim Nennerverfahren (nicht beim AKF-Verfahren) auf: b_0 lief immer gegen seinen höchsten zugelassenen Wert 3 (Fig. 31 a), im übrigen sehen die Ergebnisse gut aus. Der Mundverschluß bei /u/, vielleicht auch bei /y/, wird jedoch nicht nachgebildet (in (a) trat aus unbekanntem Grund bei /y/ Gleitkommaüberlauf auf). Wird b_0 festgehalten (b), so ändern sich die Rohrformen nicht sehr, nur bei /a/ ergibt sich ein

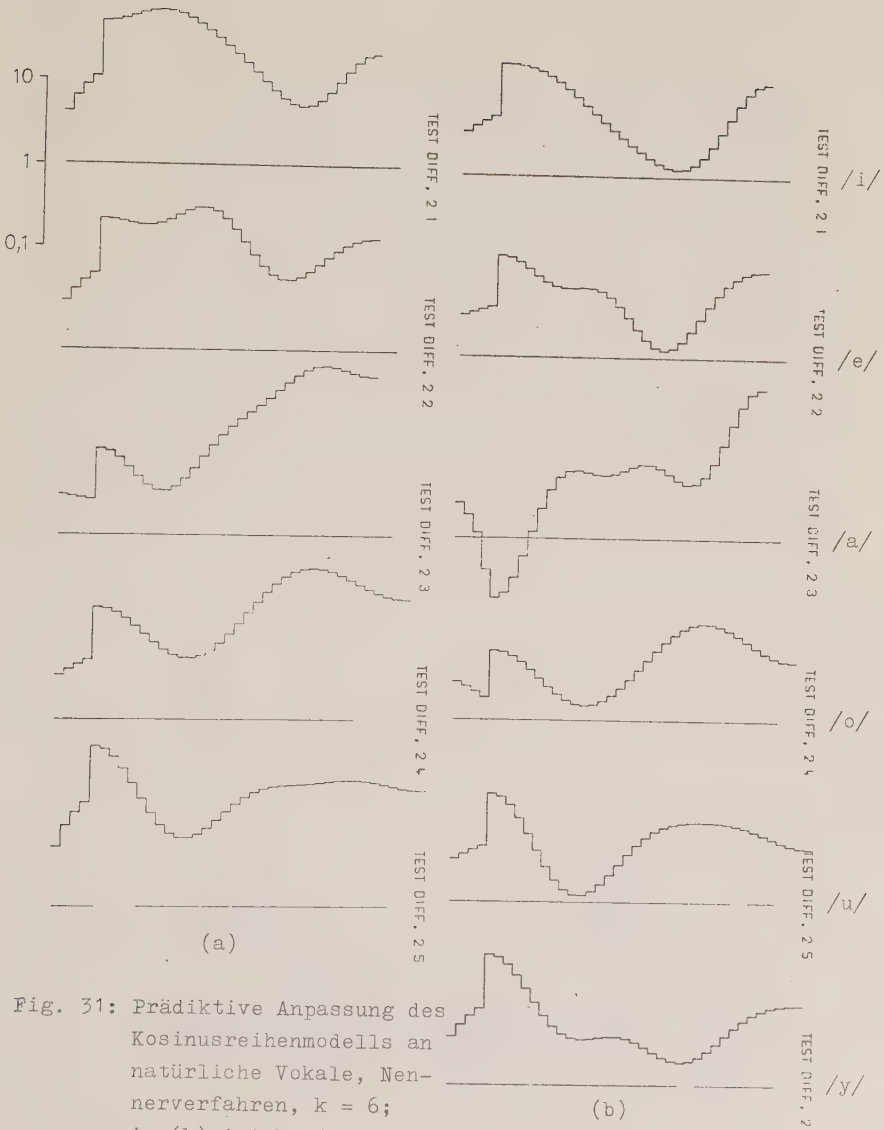


Fig. 31: Prädiktive Anpassung des Kosinusreihenmodells an natürliche Vokale, Nennerverfahren, k = 6; in (b) ist b_0 fest

Fehler (akustisch wird hier der 2. Formant 200 Hz zu hoch). Es wurde vermutet, daß die im Modell nicht berücksichtigten zusätzlichen Dämpfungsquellen des wirklichen Stimmkanals sowohl für das Weglaufen von b_0 als auch für die zu große Mundöffnung bei /u/ verantwortlich sind, da die höhere Dämpfung durch Strahlungsdämpfung simuliert werden muß. Bei /u/ liegen die Bandbreiten von F_3, F_4 bei 100-200 Hz, das Modell dagegen liefert mit Fant-/u/ nur 1 bzw. 4 Hz! Um eine grobe allgemeine Entdämpfung durchzuführen, wurde das Signal im Analyseintervall mit $e^{0,003 f_s t}$ multipliziert, d.h. alle Bandbreiten um 34 Hz verringert. Dies änderte an den Kurven praktisch nichts, die Bandbreitenverringerung war jedoch zu beobachten und die Fehlerquadratsumme war kleiner (außer bei /a,o/). Ein Faktor $e^{0,01 f_s t} \triangleq 110$ Hz hatte dagegen keine weitergehende Wirkung mehr. Eine derart grobe Entdämpfung ist wohl nicht geeignet, die Mängel des Modells auszugleichen, z.B. die hohe Dämpfung an engen Einschnürungen nachzubilden. Tab. 5 zeigt die Formantfrequenzen und die relative Fehlerquadratsumme für die Ergebnisse der drei Versuche (ohne Entdämpfung nach Fig. 31 a,b und mit der 34-Hz-Entdämpfung) im Vergleich zu den Formanten der Vokale. Diese wurden je zweimal aus den Wurzeln von Prädiktorpolynomen ermittelt; einmal liegt die AKF eines Hamminggewichteten 300-Punkte-Abschnitts zugrunde, einmal die Autokovarianzmatrix eines 20-Punkte-Abschnitts nach dem Glottisverschluß (alles auf $f_s/4 = 8825$ Hz bezogen). - Hinzunahme von b_6 ($k = 7$) gab bereits sinnlos scheinende Welligkeiten (ähnlich $k = 9$ bei Fant-Vokalen) und größere Fehler, selbst wenn die Iteration nicht von der neutralen Rohrform, sondern schon von den Koeffizienten der Kosinusreihen-Approximation der Fant-Querschnittsfunktionen startete. Ebenso lieferte das AKF-Verfahren i.a. schlechtere Ergebnisse, was auch an den Formanten und dem Fehlerquadrat abzulesen ist, und versagte oft (Fig. 32). In (b) wurde wieder eine Entdämpfung von 34 Hz versucht, indem beim Modell $H_u(z)$ durch $H_u(ze^{0,003})$ ersetzt wurde.

Die mittlere Rechenzeit pro Vokal betrug 2/3 min nach dem Nenner- und 2 min nach dem AKF-Verfahren.

	aus Prädiktor		nach Fig. 31		entdämpft
/i/	(20)	(300, Hamm.)	(a)	(b)	(34 Hz)
F_1	254	262	279	296	285
F_2	2214	2225	2164	2219	2159
F_3	2747	2755	2607	2695	2617
$10^4\Phi/\Sigma s_i^2$			963	1240	912
/e/					
F_1	337	396	356	371	369
F_2	1955	2044	1944	1963	1946
F_3	2292	2436	2290	2326	2302
$10^4\Phi/\Sigma s_i^2$			379	473	347
/a/					
F_1	806	657 917	775	807	773
F_2	1128	1179	1096	1311	1092
F_3	2598	2599	2592	2642	2573
$10^4\Phi/\Sigma s_i^2$			45	25	56
/o/					
F_1	514	474	528	538	529
F_2	912	919	909	924	909
F_3	2463	2454	2646	2681	2598
$10^4\Phi/\Sigma s_i^2$			50	82	73
/u/					
F_1	345	332	374	386	373
F_2	754	610	908	839	887
F_3	1678	?	2171	2395	2218
$10^4\Phi/\Sigma s_i^2$			208	459	175
/y/					
F_1	253	270		293	271
F_2	1440	1444	Gleit-komma-über-lauf	1466	1459
F_3	2020	2007		2095	2085
$10^4\Phi/\Sigma s_i^2$			637	1410	594

Tab. 5: Formantfrequenzen und relative Fehlerquadratsumme der Ergebnisse des Nennerverfahrens, natürliche Vokale

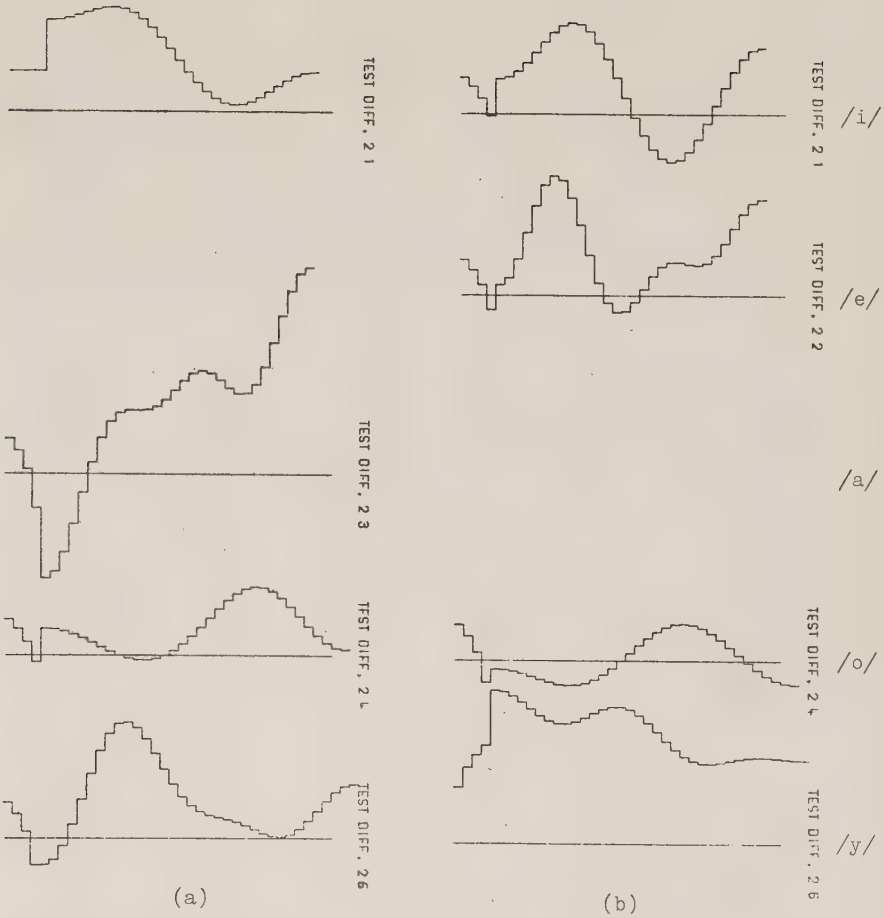


Fig. 32: Prädiktive Anpassung des Kosinusreihenmodells an natürliche Vokale, AKF-Verfahren, k = 6; in (b) wurden die Bandbreiten um 34 Hz verringert

5.4. Zusammenfassung und Diskussion

Die bisherigen Ansätze zur Bestimmung der Querschnittsfunktion des menschlichen Stimmkanals aus dem Sprachsignal für Vokale weisen schwerwiegende Nachteile auf, die die Ergebnisse unsicher machen. Die Algorithmen gehen zumeist von einfachen, stark idealisierten Annahmen über die physikalischen Eigenschaften des Stimmkanal-Rohres aus, wie Hartwandigkeit, innere Verlustfreiheit, Konzentration der Verluste an einem Ende. Oft muß die Länge des Rohres bekannt sein, ebenso die Abschlußimpedanz. Schwierigkeiten macht auch die Messung der Übertragungsfunktion (oder ihrer Äquivalente), besonders der die halbe Information enthaltenden Formantbandbreiten, und die Elimination des unbekannten Glottissignal-Einflusses. Die Einflüsse des Gesamtfrequenzganges, der Vernachlässigung der Wandimpedanz und der zugrunde liegenden physikalischen Modellvorstellungen bzgl. der Dämpfungsverteilung wurden exemplarisch an zwei Verfahren untersucht. Sie sind so groß, daß eine Anwendung solcher direkten Rechenverfahren i.a. vermutlich nicht oder nur mit großer Unsicherheit die wirkliche Querschnittsfunktion liefert, wenn sie nach neuesten Ergebnissen auch für gewisse praktische Anwendungen geeignet erscheinen (z.B.[1]).

Um nicht an sehr einfache Annahmen über den Stimmkanal gebunden zu sein und um die Menge möglicher Querschnittsfunktionen einschränken zu können, wurden diese nun durch zwei Modelle mit wenigen Parametern dargestellt, die iterativ bestimmt werden sollten. Das eine war eine Kosinusreihenentwicklung der logarithmischen Querschnittsfunktion, das andere ein kompliziertes Artikulatormodell mit bis zu fünf Parametern. In allen Fällen wurde der Stimmkanal als verlustlose Stufenleitung mit harter Glottis und einer RL-Parallelschaltung als Strahlungsimpedanz modelliert. Außer der Strahlungsdämpfung wurde, im Gegensatz zu den meisten bisherigen Verfahren, noch der bei tiefen Frequenzen recht wesentliche Einfluß der mitschwingenden Wand als RL-Serien-Querzweig am Rohreingang mit festen geschätzten Werten berücksichtigt. Die Bestimmung der Modellparameter erfolgte mit einem Marquardt-Iterationsprogramm unter Anpassung von akustischen Modellgrößen an gemessene Größen. Versuche, die logarithmischen Spektren anzupassen, scheiterten schon bei kleinen Abweichungen vom Idealfall, etwa durch den

Einfluß des Glottisspektrums. Falls Zeitintervalle existieren, in denen die Glottis geschlossen ist, der Eingangsschallfluß also verschwindet, sollte sich der Einfluß des unbekannten Glottisspektrums dadurch eliminieren lassen, daß man die Analyse auf die Glottisverschlußintervalle beschränkt. Dies wäre mit dem Autokovarianz-Prädiktionsverfahren möglich, das jedoch leicht zu Instabilitäten führt. Stattdessen wurde ein aus der Übertragungsfunktion des anzupassenden Parameter-Modells hergeleitetes Inversfilter auf das Sprachsignal angewendet und sein Ausgangssignal im Glottisverschlußintervall iterativ minimiert; eine Art Prädiktion unter Nebenbedingungen. Die Kürze des Analyseintervalls ist auch dabei sicherlich eine Quelle von Ungenauigkeiten. Ein in der Arbeit nicht beschriebener Versuch, die Analyse auf zwei aufeinanderfolgenden Verschlußintervallen auszuführen, um die Zahl der angepaßten Daten zu verdoppeln, verbesserte das Ergebnis nicht. Zur Anwendung dieses Verfahrens auf natürliche Sprache wurde eine Methode zur Bestimmung des Glottisverschlußzeitpunktes aus dem Sprachsignal entwickelt; das Ende des Verschlußintervalls ist nicht bestimmbar. Ein nicht exakter Glottisverschluß wird zu Fehlern führen können, ebenso die Verschmierung des Verschlußintervalls durch Übertragungssysteme mit nichtlinearem Phasengang bei der Sprachaufnahme.

Das Artikulatormodell wurde nur an synthetischen Vokalen getestet; es liefert oft recht gute Ergebnisse, versagt aber auch oft, je nach Vokal und Wert der beim Iterationsprozeß frei wählbaren Größen (Startwerte, Grenzfrequenz usw.), ohne daß eine Systematik zu erkennen wäre. Effekte endlicher Rechengenauigkeit scheinen eine Rolle zu spielen.

Das Kosinusreihenmodell mit angesetztem Larynxrohr führt zu besseren Ergebnissen und arbeitet sicherer, ist auch weit weniger rechen- und kernspeicherintensiv. Ein Kunstgriff, der die Konvergenz sehr beschleunigt oder überhaupt erst ermöglicht, ist das Festhalten der Kosinusreihen-Koeffizienten mit geradem Index während der ersten Iterationsschritte bei Start vom homogener Rohr. Die Anwendung auf natürliche Vokale führt zu systematischen Abweichungen: Multiplikation der Querschnitte mit einem Faktor > 1 und Unmöglichkeit, enge Lippenöffnungen herauszubekommen. Dies liegt offenbar an der nicht wirklichkeitsgetreuen Vernachlässigung der Rohrdämpfung durch Viskosität

und Wärmeleitung. Eine Überprüfung der Ergebnisse für natürliche Vokale konnte nicht durchgeführt werden, sie sehen aber vernünftig aus.

Bei beiden Modellen mußten gewisse Informationen vorgegeben werden. Das Artikulatormodell arbeitet nur einigermaßen sicher, wenn die Kieferstellung bekannt und fest ist, nach Möglichkeit auch der Gesamtmaßstab. Die Kieferstellung ließe sich relativ leicht messen. Beim Kosinusreihenmodell muß die Rohrlänge (als Vielfaches der Segmentlänge) bekannt sein. Die Länge bei wirklichen Vokalen optimal zu bestimmen, wurde mangels Vergleichsmöglichkeit der Ergebnisse mit der wirklichen Querschnittsfunktion nicht versucht. Auch von einer Variation der Wandimpedanz wurde abgesehen; hier nicht dargestellte Versuche lassen ihre prinzipielle Bestimmbarkeit zusammen mit der Querschnittsfunktion vermuten, was in der Praxis aber wohl scheitern dürfte, da die Meßgenauigkeit zu gering ist.

Das Artikulatormodell ist grundsätzlich schlecht zur Aufdeckung des Zusammenhanges zwischen Artikulatorstellungen und Querschnittsfunktion (Fig. 1) geeignet, da Annahmen hierüber bereits im Modell enthalten sind; nur deren Abänderung unter Kontrolle durch sagittale Röntgenbilder könnte hier helfen. Das Kosinusreihenmodell kann hier mehr Information liefern; die Schallausbreitung muß hierzu aber noch realistischer modelliert werden. Möglicherweise ist das Modell auch anfällig gegen Ungenauigkeiten, da es die Menge möglicher Querschnittsfunktionen (bis auf deren Glättung) nicht a priori einschränkt. - Bei der Anwendung der Verfahren auf Lautübergänge, die noch nicht versucht wurde, sollte sich die relativ lange Rechenzeit verkürzen und die Sicherheit erhöhen, da das Ergebnis einer Iteration als Startwert für die nächste, bei der sich die Rohrform nur wenig geändert hat, benutzt werden kann.

In jüngster Zeit gab es zwei neue Ansätze zur Ermittlung von Modellparametern. Jospa [70] entwickelte ein Verfahren, das ein Variationsprinzip benutzt, geht allerdings wieder von starken Idealisierungen aus und muß doch wieder iterative Methoden zur Gleichungslösung verwenden. Atal schließlich zeigte[66], daß Modellparameter als Linearkombinationen von Funktionen akustischer Größen dargestellt werden können; die Koeffizienten werden durch eine Art Lernvorgang ermittelt. Möglicherweise besteht also bei der Verwendung von Parameter-Modellen nicht

wirklich ein Zwang zur Benutzung von Iterationsverfahren, die rechenzeit- und speicheraufwendig sind.

Ich danke Herrn Prof. Dr. M. R. Schroeder für die Aufgabenstellung und für fördernde und ermutigende Gespräche; ferner ihm sowie Dr. P. B. Denes und Dr. B. S. Atal für die Ermöglichung eines Arbeitsaufenthaltes bei den Bell Telephone Laboratories, Murray Hill, N. J.
Herrn Dr. B. Wagener und Herrn W. Steinhoff habe ich für ihren Einsatz und Rat bei allen Problemen mit dem institutseigenen Rechnersystem zu danken.

Anhang

A. Lineare Prädiktion [26, 27, 38, 39, 41, 60, 71]

A.1. Grundlegendes

Es sei eine Zahlenfolge (s_k) gegeben; in der Praxis handelt es sich um Abtastwerte eines Signals $s(t)$. Dann kann man versuchen, jedes s_k durch eine Linearkombination von p vorangegangenen Werten zu approximieren:

$$\hat{s}_k = -\sum_{j=1}^{p} a_j s_{k-j} \qquad (A:1)$$

(das Minuszeichen ist willkürlich, aber praktisch). Die a_j heißen Prädiktorkoeffizienten. Mit $a_0 := 1$ läßt sich der Prädiktionsfehler darstellen als

$$e_k = s_k - \hat{s}_k = \sum_{j=0}^{p} a_j s_{k-j} \; . \qquad (A:2)$$

$$A(z) := \sum_{j=0}^{p} a_j z^{-j} \qquad (A:3)$$

heißt Prädiktorpolynom. Die a_j sind so zu bestimmen, daß das quadratische Fehlermaß

$$E_p(a_1,\ldots,a_p) := \sum_k w_k e_k^2 \quad (w_k \geq 0 \text{ Gewichte}), \qquad (A:4)$$

minimiert wird, d.h. $\partial E_p / \partial a_j = 0$, $j = 1,\ldots,p$. Definiert man die Autokovarianzmatrix

$$\Phi_{ij} := \sum_k w_k s_{k-i} s_{k-j} \; , \qquad (A:5)$$

so lassen sich die Prädiktorkoeffizienten aus den Gleichungen

$$\sum_{j=1}^{p} \Phi_{ij} a_j = -\Phi_{i0} \; , \; i = 1,\ldots,p, \qquad (A:6)$$

berechnen. Dann ist das Fehlerminimum

$$E_{p\,min} = \Phi_{00} - \sum_{i,j=1}^{p} \Phi_{ij} a_i a_j = \sum_{j=0}^{p} \Phi_{0j} a_j \; . \qquad (A:7)$$

Für nichtsinguläres $\underline{\Phi}$ löst man (A:6)vorteilhaft nach dem Choleski-Verfahren, Anhang B, wenn nicht der Fall nach A.3 vorliegt.

A.2. Das Autokovarianzverfahren

Seien jetzt die w_k in einem Intervall $K_1 \leq k \leq K_2$, und nur dort, von Null verschieden ($K_2 \to \infty$ zulässig). Dann gilt offenbar: Die Prädiktion ist genau dann exakt ($e_k = 0$) auf diesem Intervall, wenn sich das Signal s_k dort als freier Ausschwingvorgang eines Nur-Pole-Filters mit p Polen auffassen läßt; denn (A:2) hat die Gestalt einer ein solches Filter mit der Übertragungsfunktion $1/A(z)$ beschreibenden Differenzen-

gleichung mit e_k als Eingangs- und s_k als Ausgangssignal. Wenn es sich tatsächlich um einen solchen Ausschwingvorgang eines q-Pole-Filters handelt, ermöglicht die Prädiktion also die Bestimmung der Übertragungsfunktion bis auf einen Faktor, falls $p = q$ gewählt wird. Für $p > q$ wird die Kovarianzmatrix singulär, die Lösung vieldeutig; bei Abweichung von der exakten Prädizierbarkeit gilt das i.a. nicht, doch kann man dann das günstigste p dadurch bestimmen, daß sich bei Benutzung größerer p-Werte das Fehlermaß E_p nicht mehr nennenswert verringert.

Daß das Eingangssignal im untersuchten Intervall verschwindet, ist nicht erforderlich, wenn das Intervall nicht zu kurz ist. 1/A(z) approximiert dann das Gesamtspektrum; für die Bestimmung der Übertragungsfunktion muß das Anregungsspektrum also möglichst flach sein. Wird die Intervallänge groß gegen p, so verschwindet der Unterschied zum Verfahren nach A.3.

A.3. Das Autokorrelationsverfahren

Hat das Signal endliche Energie, was z.B. bei einer Impulsantwort oder einem endlichen Signalsegment (z.B. Hammingfenster) stimmt, so können alle $w_k = 1$ sein, die Autokovarianzmatrix nimmt Töplitz-Form an:

$$\Phi_{ij} = R_{i-j} \; ; \quad R_i := \sum_k s_k s_{k+i} \tag{A:8}$$

ist die Autokorrelationsfunktion, (A:6) lautet jetzt

$$\sum_{j=1}^{p} R_{k-j} a_j = -R_k \; , \; k = 1,\ldots,p. \tag{A:9}$$

Nach [60] approximiert dann 1/A(z) das Spektrum S(z) unter besonderer Berücksichtigung der Maxima. Der Prädiktionsfehler kann nicht verschwinden. Die Wahl von p erfolgt wie in A.2.

Die AKF der Impulsantwort $(h_k) = \mathcal{Z}^{-1}(1/A)$ des Prädiktorfilters erfüllt ebenfalls (A:9), aber für alle $k > 0$. Jetzt lassen sich zwei Aufgaben stellen:

1.) Ein Stück $R_0,\ldots,R_p$ einer AKF ist gegeben; konstruiere ein Filter $1/A_p(z)$, dessen Impulsantwort diese AKF hat.
2.) Die Übertragungsfunktion $1/A_p(z)$ ist gegeben; konstruiere die AKF der Impulsantwort.

Beide Aufgaben lassen sich leicht lösen, indem man einen Satz von Prädiktorpolynomen $A_n(z)$, $n = 0,\ldots,p$, $a_0^n := 1$ (n ist Index, nicht Exponent) vom Grad n betrachtet, der dem p-tupel $(R_1/R_0,\ldots,R_p/R_0)$ umkehrbar eindeutig zugeordnet ist, und zwar

soll für jedes n (A:9) gelten:

$$\sum_{j=0}^{n} R_{k-j} a_j^n = 0, \quad k = 1,\ldots,n, \quad n = 1,\ldots,p. \qquad (A:10)$$

1. Aufgabe: Wenn die Matrix (R_{k-j}), $j,k = 1,\ldots,n-1$, nichtsingulär ist, folgen die Rekursionsformeln (zu A_n^x siehe (2:22))

$$\begin{aligned} a_k^n &= a_k^{n-1} + a_n^n a_{n-k}^{n-1}, \qquad a_k^0 = \delta_{0k}, \\ \text{oder} \quad A_n(z) &= A_{n-1}(z) + a_n^n z^{-1} A_{n-1}^x(z), \quad A_0(z) = 1. \end{aligned} \qquad (A:11)$$

Multipliziert man (A:11) mit R_{n-k} und summiert von 0 bis n, so wird noch

$$a_n^n = -\Big(\sum_{k=0}^{n-1} R_{n-k} a_k^{n-1}\Big) \Big/ \Big(\sum_{k=0}^{n-1} R_k a_k^{n-1}\Big). \qquad (A:12)$$

2. Aufgabe: Wir ersetzen in (A:11) k durch n-k und eliminieren a_{n-k}^{n-1}, was die inversen Gleichungen ergibt:

$$\begin{aligned} a_k^{n-1} &= (a_k^n - a_n^n a_{n-k}^n)/(1 - (a_n^n)^2) \\ \text{oder} \quad A_{n-1} &= (A_n - a_n^n A_n^x)/(1 - (a_n^n)^2). \end{aligned} \qquad (A:13)$$

Zu dem gegebenen Polynom A_p konstruieren wir dann $A_{p-1},\ldots,A_1$, danach die AKF aus (A:10) mit $k = n = 1,\ldots,p$, $n=p$ für $k > p$, unter Vorgabe von R_0 (z.B. = 1).

I.a. muß das Prädiktorfilter $1/A_p(z)$ stabil sein, d.h. die p Wurzeln z_j von A_p müssen im Einheitskreis liegen. Wenn das so ist, gilt:

$$|a_p^p| = \prod_{j=1}^{p} |z_j| < 1. \qquad (A:14)$$

Außerdem folgt aus (A:11) oder (A:13): Wenn $|a_n^n| < 1$, so sind A_{n-1}^{-1} und A_n^{-1} entweder beide stabil oder beide instabil (Beweis mit dem Satz von Rouché, Funktionentheorie [27, 71]). Notwendig und hinreichend für Stabilität von A_p^{-1} ist also $|a_n^n| < 1$ für $n = 1,\ldots,p$. Bei der 1. Aufgabe ist das, von Rechenungenauigkeiten abgesehen, immer der Fall, weil die Matrix (R_{i-j}), $i,j = 0,\ldots,p$, positiv definit ist bzw. weil der Prädiktionsfehler größer als Null ist [60, 71].

B. Das Choleski-Verfahren [42]

Zu jeder positiv definiten symmetrischen n x n-Matrix $\underline{A} = (a_{ij})$ existiert eine untere Halbdiagonalmatrix $\underline{V} = (v_{ji})$, $v_{ji} = 0$ für $i > j$, so daß

$$\begin{aligned} \underline{A} &= \underline{V}\underline{V}^T, \\ a_{ij} &= a_{ji} = \sum_{k=1}^{i} v_{ik} v_{jk}, \quad i \le j. \end{aligned} \qquad (B:1)$$

Durch Auflösen von (B:1) nach dem Term mit k = i erhält man für die v_{ji} sofort die Rekursionsformeln

$$\left.\begin{aligned} v_{ii}^2 &= a_{ii} - \sum_{k=1}^{i-1} v_{ik}^2, \\ v_{ji} &= (a_{ij} - \sum_{k=1}^{i-1} v_{ik}v_{jk})/v_{ii}, \quad j = i+1,\ldots,n, \end{aligned}\right\} \quad i = 1,\ldots,n. \qquad (B:2)$$

Für die Determinante gilt

$$\det \underline{A} = (\det \underline{V})^2 = \prod_{i=1}^{n} v_{ii}^2. \qquad (B:3)$$

Zur Vermeidung von Gleitkomma-Überlauf berechnet man statt des Produktes (B:3) zweckmäßig gleich log det $\underline{A}$ als Summe der log v_{ii}^2.

Soll ein lineares Gleichungssystem

$$\underline{A}\,\underline{x} = \underline{b}, \qquad \underline{x},\underline{b} \in \mathbb{R}^n, \qquad (B:5)$$

gelöst werden, so liefert die Zerlegung (B:1) eine Zwischenlösung $\underline{y}$ mit

$$\underline{V}\,\underline{y} = \underline{b}, \qquad \underline{V}^T\underline{x} = \underline{y}. \qquad (B:6)$$

Zusätzlich zu (B:2) sind also die Rekursionsformeln

$$\begin{aligned} y_i &= (b_i - \sum_{k=1}^{i-1} v_{ik}y_k)/v_{ii}, \quad i = 1,\ldots,n, \\ x_i &= (y_i - \sum_{k=i+1}^{n} v_{ki}x_k)/v_{ii}, \quad i = n,n-1,\ldots,1, \end{aligned} \qquad (B:7)$$

auszurechnen. Handelt es sich bei (B:5) um die Bestimmung der Prädiktorkoeffizienten nach (A:6), so liefert nach (A:7)

$$E_{p\,min} = \Phi_{00} - \underline{x}^T\underline{b} = \Phi_{00} - \underline{y}^T\underline{y} \qquad (B:8)$$

die Quadratsumme des Prädiktionsfehlers.

Die v_{ji} mit $i < j$ lassen sich in einer Hälfte der Matrix $\underline{A}$ abspeichern. Für die v_{ii}^2 benötigt man bloß ein einziges Speicherwort, wenn nur die Determinante berechnet werden soll. Bei Lösung des Gleichungssystems (B:5) legt man die $1/v_{ii}$ im selben Speicherbereich wie $\underline{x}$ ab, $\underline{y}$ wird zusätzlich benötigt. Alle a_{ij} bleiben also erhalten, ohne daß sich weiterer Speicherbedarf ergäbe.

C. Literaturverzeichnis

1 R.G. Crichton, F. Fallside; 8th Intern. Congr. Acoust., London 1974, p. 310.

2 B.S. Atal, M.R. Schroeder; Bell Syst. Tech. J. 49 (1970) 1973-1986.

3 G. Fant: Acoustic Theory of Speech Production; Mouton, Den Haag/Paris, 2. Aufl. 1970.

4 J.M. Heinz, K.N. Stevens; 5th Intern. Congr. Acoust., Liège 1965, A44.

5 J.S. Perkell; 5th Intern. Congr. Acoust., Liège 1965, A32.

6 M. Sawashima; J. Acoust. Soc. Japan 27 (1971) 425-434.

7 F.D. Minifie, C.A. Kelsey, J.A. Zagzebski; J. Acoust. Soc. Amer. 49 (1971) 1857-1860.

8 H.S. Cooker; Speech Transmission Laboratory, Quarterly Progress and Status Report 2-3/1973 (Stockholm), pp. 1-12.

9 C.H. Coker, N. Umeda, C.P. Browman; IEEE Trans. AU-21 (1973) 293-298.

10 P. Mermelstein; J. Acoust. Soc. Amer. 53 (1973) 1070-1082.

11 J. Kacprowski; in: W. Jassem (Ed.): Speech Analysis and Synthesis, Vol. I; Warschau 1968.

12 J.L. Flanagan: Speech Analysis, Synthesis and Perception; Springer-Verlag, Berlin/Heidelberg/New York 1965.

13 J.L. Flanagan, L.L. Landgraf; IEEE Trans. AU-16 (1968) 57-64.

14 K. Ishizaka, J.L. Flanagan; Bell Syst. Tech. J. 51 (1972) 1233-1268.

15 T. Tarnóczy, K. Vicsi; 8th Intern. Congr. Acoust., London 1974, p. 231.

16 R.L. Miller; J. Acoust. Soc. Amer. 31 (1959) 667-677.

17 I.R. Titze; Dissertation April 1973, Department of Physics and Astronomy, Brigham Young University, Provo, Utah.

18 P.M. Morse: Vibration and Sound; McGraw-Hill, New York/Toronto/London, 2. Aufl. 1948.

19 E. Meyer, E.G. Neumann: Physikalische und Technische Akustik; Vieweg&Sohn, Braunschweig 1967.

20 G. Ungeheuer: Elemente einer akustischen Theorie der Vokalartikulation; Springer-Verlag, Berlin/Göttingen/Heidelberg 1962.

21 J.W. Miles; J. Acoust. Soc. Amer. 16 (1944) 14-19 und 17 (1946) 259-284.
22 M.B. Lesser, J.A. Lewis; J. Acoust. Soc. Amer. 51 (1972) 1664-1669 und 52 (1972) 1406-1410.
23 L.R. Rabiner et al.; IEEE Trans. AU-20 (1972) 322-337.
24 B. Gold, C.M. Rader: Digital Processing of Signals; McGraw-Hill, New York 1969.
25 H.W. Schüßler: Digitale Systeme zur Signalverarbeitung; Springer-Verlag, Berlin/Heidelberg/New York 1973.
26 B.S. Atal, S.L. Hanauer; J. Acoust. Soc. Amer. 50 (1971) 637-655.
27 H. Wakita; IEEE Trans. AU-21 (1973) 417-427.
28 J. van den Berg; J. Acoust. Soc. Amer. 27 (1955) 161-168 und 332-338.
29 W. Wathen-Dunn, S.B. Michaels; Ann. New York Acad. Sci. 155 (1968) 368-378.
30 G. Fant, B. Sonesson; Speech Transmission Laboratory, Quarterly Progress and Status Report 2/1964 (Stockholm), pp. 9-21.
31 O. Fujimura, J. Lindqvist; STL-QPSR 1/1974 (Stockholm), pp. 5-10.
32 G. Fant; STL-QPSR 2-3/1972 (Stockholm), pp. 28-52.
33 K. Ishizaka, J. French, J.L. Flanagan: Direct Determination of Vocal-Tract Wall Impedance; Nov. 1973 (unveröffentlicht?).
34 M.M. Sondhi; J. Acoust. Soc. Amer. 55 (1974) 1070-1075.
35 C.M. Rader, B. Gold; Proc. IEEE 55 (1967) 149-171.
36 H.W. Strube: Sampled-Data Representation of a Non-Uniform Lossless Tube of Continuously Variable Length; J. Acoust. Soc. Amer., im Druck.
37 T. Takasugi, J. Suzuki; J. Radio Res. Labs. (Japan) 17 (1970) 153-168.
38 J.D. Markel; IEEE Trans. AU-20 (1972) 129-137 und 367-377.
39 J.D. Markel; IEEE Tranc. AU-21 (1973) 154-160.
40 J.N. Maksym; IEEE Trans. AU-21 (1973) 149-154.
41 A.N. Sobakin; Soviet Physics - Acoustics 18 (1972) 84-90.
42 R.A. Buckingham: Numerical Methods; Pitsman&Sons, London 1962.
43 M.R. Wohlers; IEEE Trans. CT-13 (1966) 356-363 und CT-14 (1967) 442-444.
44 O.N. Litvinenko, V.I. Soshnikov; Telecom. & Radio Eng. Part 2, Radio Eng. 17 (1962) Nr.9, 14-22

45 B. Gopinath, M.M. Sondhi; Bell Syst. Tech. J. 49 (1970) 1195-1214.
46 M.M. Sondhi, B. Gopinath; J. Acoust. Soc. Amer. 49 (1971) 1867-1873.
47 B.S. Atal: Determination of the Vocal-Tract Shape Directly from the Speech Wave; 1970, Veröffentlichung unbekannt.
48 A. Paige, V.W. Zue; IEEE Trans. AU-18 (1970) 7-18.
49 V.Ya. Maleyev; Radio Eng. & Elec. Phys. 7 (1962) 1692-1695.
50 H. Fujisaki, S. Obata, R. Tazaki; Ann. Rept., Eng. Res. Inst., Fac. Eng., Univ. Tokyo 30 (1971) 81-88.
51 M.R. Schroeder; J. Acoust. Soc. Amer. 41 (1967) 1002-1010.
52 P. Mermelstein; J. Acoust. Soc. Amer. 41 (1967) 1283-1294.
53 P. Mermelstein, M.R. Schroeder; 5th Intern. Congr. Acoust., Liège 1965, A24.
54 P. Mermelstein, J.E. West: The Impedance Tube Method for Acoustic Measurement of Area Functions of Non-Uniform Tubes; Juni 1968, Veröffentlichung unbekannt.
55 A. Paige, V.W. Zue; IEEE Trans. AU-18 (1970) 268-270.
56 E. Meyer, D. Guicking: Schwingungslehre; Vieweg&Sohn, Braunschweig 1974.
57 E.N. Pinson; J. Acoust. Soc. Amer. 35 (1963) 1264-1273.
58 A.V. Oppenheim, R.W. Schafer; IEEE Trans. AU-16 (1968) 221-226.
59 A.M. Noll; J. Acoust. Soc. Amer. 41 (1967) 293-309.
60 J.I. Makhoul, J.J. Wolf; Bolt Beranek and Newman Report No. 2304, Aug. 1972.
61 E.N. Pinson; 5th Intern. Congr. Acoust., Liège 1965, A37.
62 K.R. Fellbaum, H. Mangold (AEG-Telefunken FI Ulm), persönliche Mitteilung.
63 T. Nakajima et al.; 8th Intern. Congr. Acoust., London 1974, p. 230; und Speech Communication Seminar, Stockholm 1974, Vol. 1, pp. 11-20.
64 D.W. Marquardt; J. Soc. Indust. Appl. Math. 11 (1963) 431-447.
65 K.N. Stevens, A.S. House; J. Acoust. Soc. Amer. 27 (1955) 484-493.
66 B.S. Atal; 8th Intern. Congr. Acoust., London 1974, p. 234; und Speech Communication Seminar, Stockholm 1974, Vol. 1, pp. 1-9.

67 B.E.F. Lindblom, J.E.F. Sundberg; J. Acoust. Soc. Amer. 50 (1971) 1166-1179.

68 G-AE Subcommittee on Measurement Concepts; IEEE Trans. AU-15 (1967) 45-55.

69 R.C. Singleton; IEEE Trans. AU-17 (1969) 93-103.

70 P. Jospa; 8th Intern. Congr. Acoust., London 1974, p. 226; und Speech Communication Seminar, Stockholm 1974, nachgereichter Sonderdruck.

71 J.D. Markel, A.H. Gray, Jr.; IEEE Trans. AU-21 (1973) 69-79.

Lebenslauf

Ich wurde am 8. 12. 1945 in Grömitz (Holstein) geboren. Meine Eltern waren Ernst Strube, damals Abteilungsleiter für Tiefbau und später Betriebsleiter in einer chemischen Reinigung, und seine Frau Hildegard, geb. Hannig, die 1959 verstarb. Seit 1946 lebte ich in Northeim, wo ich 1952-56 die Volksschule und dann das Gymnasium (mathematisch-naturwissenschaftlicher Zweig) besuchte. 1965 bestand ich das Abitur und studiere seitdem in Göttingen Physik. Im Mai 1967 legte ich die Vordiplomprüfung ab und arbeitete ab Juli 1968 im III. Physikalischen Institut bei Prof. Kuttruff an einer experimentellen Diplomarbeit über Kavitationskeime an der Grenzfläche Quarz-Wasser. Daneben entstand Ende 1968 - Anfang 1971 eine theoretische Arbeit über nichtsphärische Blasenschwingungen. Im Februar 1970 bestand ich die Diplomprüfung und bin seitdem Doktorand bei Prof. M.R. Schroeder im III. Phys. Inst. 1970 arbeitete ich im Zusammenhang damit für drei Monate bei den Bell Telephone Laboratories in den USA.